Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 105-111 recibido 26-04-10 revisado 18-05-10 aceptado 20-05-10

ClInt: a Bilingual Spanish-Catalan Spoken Corpus of Clinical Interviews

ClInt: un corpus oral bilingüe español-catalán de entrevistas clínicas

Marta Vila, Santiago González, M. Antònia Martí
CLiC – Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
08007 Barcelona
{marta.vila, santiago.gonzalez, amarti}@ub.edu

Joaquim Llisterri, María Jesús Machuca
Grup de Fonètica – Departament de Filologia Espanyola
Universitat Autònoma de Barcelona
Edifici B, 08193 Bellaterra, Barcelona
{joaquim.llisterri, mariaJesus.Machuca}@uab.cat

Resumen: En este artículo se presenta ClInt (*Clinical Interview*), un corpus oral bilingüe español-catalán que contiene un total de 15 horas de entrevistas clínicas. Está formado por archivos sonoros alineados con transcripciones a varios niveles que comprenden información ortográfica, fonética y morfológica, además de codificación lingüística y extralingüística. Se trata de un recurso hasta el momento inexistente para estas lenguas que ofrece múltiples posibilidades de explotación desde una amplia variedad de disciplinas, tanto las vinculadas a la Lingüística como las que se relacionan con el Procesamiento del Lenguaje Natural.
Palabras clave: Corpus oral, corpus bilingüe, entrevista clínica.

Abstract: In this paper we present ClInt (Clinical Interview), a bilingual Spanish-Catalan spoken corpus that contains 15 hours of clinical interviews. It consists of audio files aligned with multiple-level transcriptions comprising orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding. This is a previously non-existent resource for these languages and it offers a wide-ranging exploitation potential in a broad variety of disciplines such as Linguistics, Natural Language Processing and related fields.
Keywords: Spoken corpus, bilingual corpus, clinical interview.

1 Introduction

Corpus availability has become indispensable for performing studies in many scientific fields. Nowadays, these language resources are fundamental in disciplines such as Linguistics, Natural Language Processing (NLP) and related fields.

Spoken corpora are those most in demand, probably due to their shortage and the difficulty involved in obtaining them, not only in the transcription procedure, but also in the recording. In this sense, one of the most valuable types is the one that captures real —not artificially elicited— communicative situations. Spoken corpora in professional situations are especially difficult to obtain, because it is not easy to gain access to certain environments, such as trials, business meetings, or clinical interviews.

In this paper we present ClInt (Clinical Interview),[1] a bilingual Spanish-Catalan spoken corpus of clinical interviews, a hitherto non-existent resource for these languages. It consists of audio files aligned with multiple-level transcriptions containing orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding.

The remainder of this paper is structured as follows: in Section 2, we present the related work done in this area. In Section 3, we provide an overview of the corpus. Section 4 is devoted

[1] The corpus and source URLs mentioned in this paper appear in the appendix.

ISSN 1135-5948

to corpus development. In Section 5, future research possibilities are suggested. Finally, Section 6 sets out some final remarks about this project.

2 Related Work

To the best of our knowledge, ClInt is the first bilingual Spanish-Catalan spoken corpus of clinical interviews. Moreover, there are very few corpora of this type in other languages. The DiK-corpus is particularly relevant in this sense. It consists of the transcriptions of 25 hours of audio recordings of monolingual and interpreted doctor-patient communication in German, Turkish, Portuguese and Spanish.

Despite the shortage of clinical interview corpora, in more general terms, there do exist spoken conversational corpora, both in Spanish and Catalan. In Spanish, some examples are CORLEC (*Corpus Oral de Referencia de la Lengua Española Contemporánea*[2]) (Marcos, 1991), the *Corpus de conversaciones coloquiales*[3] (Briz, 2001) and the spoken section in CREA (*Corpus de Referencia del Español Actual*[4]) (RAE). Our major reference in Catalan is COC (*Corpus Oral de Conversa Coloquial*[5]) (Payrató and Alturo, 2002), contained in the CCCUB (*Corpus del Català Contemporani de la Universitat de Barcelona*[6]).

Moreover, there exist corpora including speech by sick and disabled people, and by people with language disorders (Peraita and Grasso, 2009; Navarro and San Martín, 2009). Also, recorded clinical interview simulations for doctor training can be found (Borrell, 2000).

Finally, there has been some work in the literature with regard to clinical therapist skills training in virtual environments. In this context, the patient is a virtual human and the doctor has to interact with this virtual human in order to improve his skills in the process (Kenny et al., 2007; 2008).

3 Corpus Overview

The corpus is comprised of a total of 15 hours of recordings divided into 40 clinical interviews of an average of 22 min each. These interviews correspond to four different residents (ten interviews for each resident).

The recordings were carried out in the pneumology clinic of a hospital in the Barcelona metropolitan area. Catalonia is a bilingual community where Catalan and Spanish coexist. As the recordings were made giving absolute freedom to participants with respect to their language usage, this bilingualism is reflected in the corpus. Furthermore, the corpus displays Spanish and Catalan dialectal variants.

The ClInt corpus consists of the audio files aligned with their orthographic transcriptions (with linguistic and extralinguistic encoding), their phonetic transcriptions, as well as their morphosyntactic analysis. All this information is stored in a database.

4 Corpus Development

The ClInt corpus (Figure 1) was recorded using a stereo digital recorder (SANYO, ICR-RS176NX) and a uni-directional condenser microphone (FoneStar, BM-704BL). The characteristics of this equipment ensure that the corpus is available for further phonetic studies. These recordings were manually transcribed using conventional spelling and encoded in XML format using the Transcriber (Barras et al., 2001), a tool for assisting in the manual transcription and encoding of speech signals that provides a user-friendly interface. This tool allows for the alignment between the audio and the transcription.

The basic unit of the text in the corpus is the 'breath group',[7] understood as a discourse stretch of speech between pauses (a pause is defined as a period of silence between 200 and 500 ms). Breath groups can be full (with speech uttered), empty (pauses above 500 ms) or with overlapping (when two people speak at the same time). A breath group generally corresponds to a register in the database and it is the unit of alignment, i.e, the audio files and the different transcription levels are synchronized at the level of breath groups.

From the manual transcription, called the Base Transcription (BT), an Orthographic Transcription (OT) and an Enriched Orthographic Transcription (EOT) were automatically obtained. The raw OT was used

[2] Reference Corpus of Contemporary Spoken Spanish

[3] Corpus of Colloquial Conversations

[4] Reference Corpus of Current Spanish

[5] Spoken Corpus of Colloquial Conversation

[6] Corpus of Contemporary Catalan of the University of Barcelona

[7] Also called 'phonic group' in the Spanish tradition.

in turn for the generation of the Phonetic Transcription (PhT) and as input for the Morphological Analysis (MA).

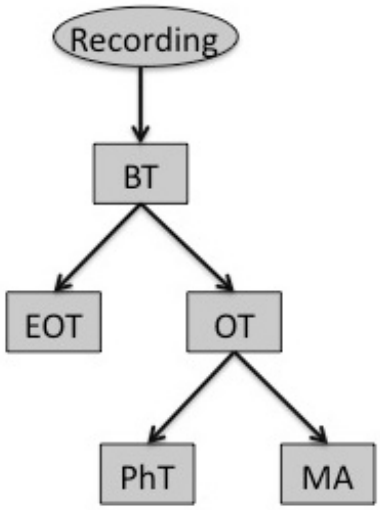

Figure 1: Corpus development scheme

4.1 Base Transcription

The BT (Figure 2) consists of a manual orthographic transcription and encoding of the audio files in XML format. For this purpose, we developed annotation guidelines and carried out a training process for all the annotators in order to avoid incoherencies in the transcription.

The orthographic transcription guidelines follow EAGLES - Expert Advisory Group on Language Engineering Standard (1996) recommendations. EAGLES general philosophy is always to use prescriptive forms and to document all the cases where this is not possible. Following these recommendations, the annotators used, whenever possible, the orthographic forms that appear in the Spanish and Catalan prescriptive dictionaries. However, with the aim of being faithful to the speakers' pronunciation, some non-prescriptive words (i.e. some onomatopoeias, interferences, unknown and mispronounced words, and abbreviated forms) were maintained and tagged. All of them are collected in a document accompanying the transcription. Numbers, acronyms and spelled words are represented as the speakers pronounce them, i.e., using the orthographically complete form. Prosodic tags are used instead of punctuation marks to ensure the correct interpretation of the text and, at the same time, to accurately reflect the spoken nature of the corpus.

The encoding is intended to be as general and scalable as possible in order to ensure the widest possible exploitation potential for ClInt. Below we list the tags corresponding to the information and phenomena that are encoded in the BT. For the sake of simplicity, we classify them into groups according to the type of information encoded.

Recording and transcription files (information about every recording and transcription file in the corpus): recording identification and date, person responsible for transcription, and transcription date.

Speakers (information about the speakers participating in the interaction): speakers' identification and sex, languages in which they are competent, and the language they (generally) use in the interview.

All the languages in which each speaker is (not) competent have a code (from 0 to 3) indicating the level of competence:

-The speaker does not understand the language.

-The speaker is able to understand the language, but is not able to speak it.

-The speaker is able to speak the language, but with certain limitations.

- The speaker is perfectly able to speak the language.

All the information related to languages is extracted from the recordings themselves. Information that is not specified or deductible from the recordings does not appear, since it is considered to be subjective.

Discourse interaction-related phenomena (information about turn-taking): turn-taking, overlaps, pauses above 500 ms.

Lexical and semi-lexical phenomena:

-Named entities: people, medicines and active principles.

-Acronyms: word formed from the initial letters of other words (e.g. *TAC* for *Tomografía Axial Computarizada*, 'computed tomography' in Spanish) [8].

-Spelled words: words uttered naming the letters that form them (e.g., *a-a-ese* for *AAS,* in this example, the patient is trying to spell the name of a medicine).

-Syllabification: words uttered separating the syllables that form them (e.g., *se-tan-ta-dos* for *setanta-dos*, 'seventy-two' in Catalan).

[8] For the sake of simplicity, we do not exemplify these phenomena using the XML tags.

-Onomatopoeias: words that reproduce the sound associated with what is named (e.g., *bumbum,* in this example, the patient is trying to reproduce the sound of fast walking).

-Interjections: words used for expressing the speaker's attitude (e.g., *ai*, in this example, the speaker is expressing pain) or for maintaining the communication between speakers (e.g., *ahà,* in this example, the speaker is communicating that he is following the conversation), among other uses.

-Abbreviated forms: words that have lost a sound or sounds at the end (e.g., *químio* for *quimioterapia*, 'chemotherapy' in Spanish).

-Mispronounced words: words that are uttered in the wrong way (e.g., *otroscopia* for *artroscopia*, 'arthroscopy' in Spanish).

-Truncated words: words that have been truncated in the interview for different reasons such as an interruption by another speaker (e.g., *magat* for *magatzem*, 'warehouse' in Catalan).

-Emphasis: words uttered prominently.

-Long sounds: lengthened sounds in a word.

-Non-understandable snippets: incomprehensible fragments.

-Unknown words: words that can be partially understood. The tag indicates that the interpretation is a guess.

-Voiced pauses: pauses in the speech in which the speaker produces a semi-lexical sound (e.g., *eee*).

Non-lexical phenomena: human and non-human noises (e.g., laughing, slams, typing).

Code-related phenomena: mixing and code switching.

Prosodic phenomena: terminal and truncated tones, following Payrató and Fitó (2008).

<Turn speaker="spk4" startTime="702.244" endTime="705.062">
<Sync time="702.244"/>
y cuando haces
<Event desc="voiced_pause" type="lexical" extent="begin"/>mmm<Event desc="voiced_pause" type="lexical" extent="end"/>
ejercicio
<Event desc="noise" type="noise" extent="begin"/>
<Event desc="long" type="pronounce" extent="begin"/>**s**<Event desc="long" type="pronounce" extent="end"/>**ientes**
<Event desc="noise" type="noise" extent="end"/>
que te falta un poco el aire<Pro desc="asc"/>
<Turn/>
<Turn speaker="spk2" startTime="705.062" endTime="706.059">
<Sync time="705.062"/>
sí el aire<Pro desc="desc"/>
<Turn/>

Figure 2: Example of Base Transcription[9]

4.2 Enriched Orthographic Transcription

The EOT (Figure 3) was automatically obtained from the BT just by changing the XML tags for more readable marks, e.g., <Turn speaker="spk4"> in Figure 2 has been changed to "Doctor" in Figure 3; or <Event desc="noise" type="noise" extent="end"/> in Figure 2 has been changed to [-noise] in Figure 3. This makes the transcription more readable.

Doctor y cuando haces <mmm> ejercicio [noise-] s:ientes [-noise] que te falta un poco el aire/
Patient sí el aire\

Figure 3: Example of Enriched Orthographic Transcription

4.3 Orthographic Transcription

The OT (Figure 4) was automatically obtained from the BT by eliminating all XML tags. Moreover, truncated words were reconstructed when they could be inferred from the context. When they could not, they were eliminated. Voiced pauses were not included either.

The OT has a neutral intermediate format suitable for automatically deriving the PhT and for carrying out the MA.

Doctor y cuando haces ejercicio sientes que te falta un poco el aire
Patient sí el aire

Figure 4: Example of Orthographic Transcription

4.4 Phonetic Transcription

The PhT is derived from the OT using SAGA (Moreno and Mariño, 1998), an automatic

[9] And when you do exercice, you feel you are breathless / Yes, I do.

Spanish phonetic transcriber, for the fragments in Spanish (Figure 5), and SEGRE (Pachès et al., 2000), an automatic Catalan phonetic transcriber, for the fragments in Catalan (Figure 6). The phonetic alphabet used in both cases is SAMPA. Although both SAGA and SEGRE take into account contextual phonetic phenomena (both inter and intra word), SEGRE considers resyllabification phenomena corresponding to spontaneous speech, e.g., the transcription [s 'i | **e** | **l** 'a j | r e] in Figure 6 considers resyllabification (in bold), while the transcription corresponding to SAGA [s 'i / **e l** / 'a j - r e] in Figure 5 does not.

Doctor i / k w a n - d o / 'a - T e s / e - x e r - T 'i - T j o / s j 'e n - t e s / k e / t e / f 'a l - t a / 'u m / p 'o - k o / e l / 'a j - r e
Patient s 'i / e l / 'a j - r e

Figure 5: Example of phonetic transcription using SAGA

Doctor i | k w a n | d o 'a | T e | s e | x e r | T 'i | T j o | s j 'e n | t e s | k e | t e | f 'a l | t a 'u m | p 'o | k o e | l'a j | r e
Patient s 'i | e | l 'a j | r e

Figure 6: Example of phonetic transcription using SEGRE[10]

4.5 Morphosyntactic Analysis

The MA (Table 1) is derived from the OT using FreeLing toolbox (Atserias et al., 2006). The MA is not strictly speaking a transcription, but a morphosyntactic analysis of all words in the corpus. A lemma and a category are assigned to every word in the corpus. Because of the spoken and sometimes non-prescriptive nature of the corpus, some questions were not held by the analyzer correctly. Thus, a manual revision of the morphosyntactic analysis had to be carried out.

Lemma	Word	Code
y	y	cc
cuando	cuando	cs
hacer	haces	vmip2s0
ejercicio	ejercicio	ncms000
sentir	sientes	vmip2s0
que	que	cs
tú	te	pp2cs000
faltar	falta	vmip3s0
el	el	da0ms0
aire	aire	ncms000
sí	sí	rg
el	el	da0ms0
aire	aire	ncms000

Table 1: Example of Morphosyntactic Analysis

5 Research Exploitation and Future Work

This corpus opens up a wide variety of possibilities in research. We want to emphasize the relevance of this corpus to disciplines such as Linguistics and NLP. Three main lines of research are being carried out. Firstly, ClInt constitutes part of a wider project, Text-Knowledge 2.0, aimed at studying language use. For this project we are developing several Catalan and Spanish corpora representative of different communicative situations. Our hypothesis is that there are fundamental differences between how linguistic structure is postulated on the basis of imagined configurations, and how it is actually expressed in live conversational contexts. More specifically, we want to identify memory storage units, that is, the way in which language is broken down into chunks based on the frequency of items and strings of items (Bybee and Hopper, 2001; Bybee, 2010).

Secondly, this corpus is especially relevant for the study of paraphrasing occurring over different registers. On many occasions, during the clinical interview, the same information is uttered by the doctor and the patient. However, in general terms, whereas doctors talk objectively using a technical register conferred by their medical knowledge and experience, patients talk subjectively, expressing their personal experience of illness, due to their lack of medical knowledge.

Thirdly, from a phonetic point of view, this type of corpora corresponds to spontaneous speech providing physical evidence of how we actually speak. Disfluencies can be studied in order to analyze how speakers plan their speech and which planning problems there are when someone says something in real conversation (Clark and Wasow, 1998). Moreover, modeling variation in spontaneous speech is also important to improve speech recognition systems. According to Nakamura, Iwano and Furui (2007) recognition performance

[10] Although SEGRE only works for Catalan, we have used the same snippet in Spanish in order to facilitate the comparison.

drastically decreases for spontaneous speech, so a paradigm shift from speech recognition to understanding is required when underlying messages of the speaker are extracted.

Finally, clinical-interview corpora are indispensable in the medical communication field. Many experts point out that doctor-patient communication has been given little attention (Clèries, 2006). Nowadays, doctors are more encouraged to perform therapeutic procedures than to talk to the patient, although on many occasions the diagnosis may be obtained solely through communication. According to some experts, the problem is that young doctors are not sufficiently trained and cover up their lack of experience with technique. Hence, many point out the need for communicative skills training in Medical Schools. Clinical-interview corpora are indispensable for doing research in this area and as real material to work with in communicative-skills training courses.

6 Final Remarks

In this paper, we have presented ClInt, a corpus of 15 hours of clinical interviews. It consists of audio files aligned with multiple-level transcriptions containing orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding. The encoding is intended to be as general and scalable as possible, as ClInt's exploitation potential is very wide-ranging. We have shown the linguistic richness of this resource, partly due to its bilingual nature. We have also described the interest of this corpus from the Linguistics and NLP perspectives, as well as from a medical point of view.

Acknowledgments

We are very grateful to Gustavo Tolchinsky, the doctor responsible for the recordings, as well as to the annotators: Alba Vindel and Esther Arias. The construction of the ClInt corpus would not have been possible without their collaboration. This work is supported by the FPU Grant AP2008-02185 from the Spanish Ministry of Education, and the Text-Knowledge 2.0 (TIN2009-13391-C04-04) and ClInt (FFI2009-06252-E/FILO) projects.

Bibliografía

Atserias, J., B. Casas, E. Comelles, M. González, Ll. Padró and M. Padró. 2006. FreeLing 1.3: Syntactic and Semantic Services in an Open-source NLP Library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*, pages 18-25, Genoa.

Barras, C., E. Geo, Z. Wu and M. Liberman. 2001. Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. *Speech Communication,* 33:5-22.

Borrell, F. 2000. *Entrevista clínica. Manual de estrategias prácticas*. SemFYC, Barcelona.

Briz, A. (Coord.) 2001. Corpus de conversaciones coloquiales. Appendix 1 in *Oralia*. ArcoLibros, Madrid.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Bybee, J. and P. Hopper. 2001. *Frequency and the Emergency of Linguistic Structure*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Clark, H. and T. Wasow. 1998. Repeating Words in Spontaneous Speech, *Cognitive Psychology*, 37:201-242.

Clèries, X. 2006. *La comunicación. Una competencia esencial para los profesionales de la salud*. Elsevier-Masson, Barcelona.

EAGLES. 1996. *Preliminary Recommendations on Spoken Texts*. EAGLES Document EAG-TCWG-STP/P, May 1996.

Kenny, P., Th. Parsons, J. Gratch, A. Leuski and A. Rizzo. 2007. Virtual Patients for Clinical Therapists Skills Training. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, *IVA,* pages 197-210, Paris.

Kenny, P., Th. Parsons, J. Gratch and A. Rizzo. 2008. Evaluation of Justina: A Virtual Patient with PTSD. In *Proceedings of 8th International Conference on Intelligent Virtual Agents, IVA,* pages 394-408, Tokio.

Marcos, F. 1991. Corpus lingüístico de referencia de la lengua española. *Boletín de la Academia Argentina de las Letras* 56: 129-155.

Moreno, A. and J. B. Mariño. 1998. Spanish Dialects: Phonetic Transcription. In *Proceedings of the 5th International*

Conference on Spoken Language Processing, ICSLP, pages 189-192, Sydney.

Nakamura, M., K. Iwano and S. Furui. 2008. Differences Between Acoustic Characteristics of Spontaneous and Read Speech and Their Effects on Speech Recognition Performance. *Computer Speech & Language*, 22(2):171-184.

Navarro, M. I. and C. San Martín. 2009. Estudio comparativo de las habilidades metalingüísticas de un niño con Trastorno Específico del Lenguaje basado en un corpus de niños con este trastorno y niños que siguen la pauta estándar de desarrollo procedentes de un corpus de niños normohablantes. In *Proceedings of the 1st International Conference on Corpus Linguistics, CILC,* pages 326-344, Murcia.

Pachès, P., C. Mota, M. Riera, M. P. Perea, A. Febrer, M. Estruch, J. M. Garrido, M. J. Machuca, A. Ríos, J. Llisterri, I. Esquerra, J. Hernando, J. Padrell, C. Nadeu. 2000. Segre: An Automatic Tool for Grapheme-to-allophone Transcription in Catalan. In *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and strategic priorities, 2nd International Conference on Language Resources and Evaluation, LREC,* pages 52-61, Athens.

Payrató, Ll. and N. Alturo (Eds.). 2002. *Corpus oral de conversa col·loquial. Materials de treball*, Publicacions de la Universitat de Barcelona, Barcelona.

Payrató, Ll. and J. Fitó. 2008. *Corpus audiovisual plurilingüe*. Publicacions de la Universitat de Barcelona, Barcelona.

Peraita, H. and L. Grasso. 2009. Corpus lingüístico de definiciones de categorías semánticas de sujetos ancianos sanos y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina. In *Proceedings of the 1st International Conference on Corpus Linguistics*, *CILC,* pages 78-88, Murcia.

A Appendix 1: Corpus and Source URLs

CCCUB corpus
<http://www.ub.edu/cccub/>

ClInt corpus
< http://clic.ub.edu/en/clint-en>

CORLEC corpus
<http://www.lllf.uam.es/ESP/Corlec.html>

Corpus de conversaciones coloquiales
<http://www.valesco.es/>

CREA corpus
<http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/D55F5BFB05D63980C1257164003F02E5?OpenDocument&i=2>

DiK corpus
<http://www1.uni-hamburg.de/exmaralda/files/k2-korpus/index.html>

EAGLES standard
<http://www.ilc.pi.cnr.it/EAGLES96/spoken tx/spokentx.html>

FreeLing
<http://www.lsi.upc.es/~nlp/freeling/>

SAGA
<http://www.talp.cat/talp/index.php/ca/recursos/eines/saga>

SEGRE
<http://www.talp.cat/Joomla_1.5.7_nou/index.php/ca/recursos/eines/segre>

Transcriber
<http://trans.sourceforge.net/en/presentation.php>