Kjær Jensen

ENTREVIS – a Spanish machine-readable text corpus

1. Short presentation

In order to conduct grammatical studies in modern Spanish, I began building a machine-readable text corpus, ENTREVIS, in early 1990. This corpus, which in some respects differs from other existing corpora[1], will contain all interviews with Spaniards published during 1990 in the two weekly Madrid magazines, *Tiempo* and *Cambio16*. The corpus contained a total of some 400,000 running words by early May 1991, when half of the corpus had been scanned in (i.e. all issues of the magazines published from 1 January 1990 to 30 June 1990).

The corpus is available in a machine-readable form (ASCII-files) and as a paper copy.

The corpus has been built to assist lexical, morphological, syntactic, semantic, textlinguistic and certain pragmatic studies[2].

Further projects: when completely compiled at the end of 1991, the whole corpus (1 January 1990 — 31 December 1990) will contain some 800,000 running words. I have planned to construct a database to allow searches, e.g. on the person interviewed, the journalist, the "theme" of the interview, and, of course, the main text (i.e. the questions and answers). Another corpus containing interviews with Spanish speaking people from Latin-America is being contemplated.

---

[1] The following list of Spanish text corpora in Scandinavia does not pretend to be complete:
PE77: (Banco de Datos de Prensa Española 1977, PE77); 2 million current words fram Spanish newspapers from June-December 1977 and an alfabetical concordance on 172 microfiches. Institutionen för Romanska Språk / Spanska, Lundgrensgatan 7, 412 56 Göteborg, Sweden.
JUR–korpus: 1291 texts pertaining to civil procedure. 402,000 Spanish words. Sampling period: 1970 ff. Compiled 1985-87 at The Copenhagen Business School, Dalgas Have, Copenhagen F, Denmark.
Biotec.ES: Spanish corpus of biotechnology. 1 million words compiled 1989 at The Copenhagen Business School, The Aarhus School of Business and The Southern Denmark Business School. All texts concern genetic engineering and comprise a maximum of 5,000 words each.

[2] For instance, the use of "tù" and "usted" must be defined as belonging to pragmatics; the use depends, among other things, on age, authority and sex of the persons involved in the dialogue. At least in one case in these interviews, the journalist is saying "usted", but the lady who is interviewed is saying "tú".

2. Composition of the corpus

2.1 The texts

The texts, which have been scanned in, include the title of the interview, in most cases a few lines about the person being interviewed or the theme, the names of the journalist and the photographer (photos only exist in the papercopy), and finally the questions and answers.

The corpus only includes full interviews, i.e. series of questions and answers involving two or three persons, excluding replies that are not part of an interview proper. Apart from transcriptions of a few police wiretaps, the interviews are intended for the general public, and the texts include interviews on many topics such as music, economics, industrial relations, etc.

2.2. The persons involved in the interviews.

The interviewer's or journalist's name is usually given before the text proper.

Only such interviews are included that present the person interviewed as a Spaniard. The majority of the individuals interviewed are living somewhere in Spain. Some are living in Castile and some, but not all, have been living their whole life in Spain. Thus, e.g. the following two interviews are both included in the corpus: a famous writer born in Galicia spending long periods of his life in Madrid or in Mallorca, and a famous scientist born in Asturias, now resident of Madrid, who spent most of his professional life in the U.S. Inversely, interviews in which the magazine identifies the interviewed individual as a Latin-American (e.g. a politician from Costa Rica) or a foreigner (e.g. Margaret Thatcher), are not included in the corpus.

The individuals interviewed are of both sexes and all ages, although children have not been interviewed.

Social groups. Interviews have mainly been conducted with individuals from the upper strata of the Spanish society, including professions such as rock singers, opera singers, dancers, musicians, actors, artists, sportsmen, scientists, authors, politicians, students, executives, etc. The corpus only includes a single interview with an illiterate individual from the lowest social strata.

3. Some observations of importance for the use of the corpus

3.1 The corpus and the concept of language.

In HERMES 6, Ole Lauridsen, Theis Riiber and Henning Søndergaard

discussed certain problems related to text corpora. In order to avoid repetitions, I refer to this article concerning the representativeness of a corpus, the relation between a corpus and the concept of language, tagging, etc.

3.2 Are these interviews true transcriptions of authentic spoken language?

This problem has been addressed by Milan Kundera who in 1985 decided to stop giving any more interviews. He motivated his refusal to the German newspaper, Die Zeit, in the following way: 'Firstly, the journalist asks us only such questions that are interesting for him, not for us. Secondly, he will use only those of our answers that suit his purpose. Thirdly, he will reformulate our answers into his own words, into his own way of thinking' (my translation[3]). According to Milan Kundera, interviews are texts produced by the journalists that may have nothing to do with the oral expression of the person interviewed. One may agree with Milan Kundera to a variable extent, but compared to the oral medium, the written medium does impose a far greater explicitness: the careful and precise composition of a sentence of the latter as opposed to the casual expression supported by gesture of the former. Many of the devices used to transmit language by speech (stress, rhythm, intonation, etc) cannot be embodied in the relatively limited repertoire of conventional orthography. So we have to admit that the interviews published in magazines cannot be considered true transcriptions of spoken language. Interviews are written texts that have certain features in common with spoken language. Roughly, they could be characterized as belonging to the continuum between spoken and written language proper. In the interviews we do not find many features typical of spontaneously spoken language (interruptions, exclamations, questions asked by both parties, etc.) or words that are extremely common in spoken Spanish but usually avoided in print. Nor will we find a highly specialized vocabulary and the complex sentence structures typical of some types of written language (for instance departmental style).

In interviews the editor's policy will be to use commonly known words and expressions (including e.g. words that are commonly used in spoken Spanish without being yet officially accepted; such words are written in italics). The editor will adopt a style that reflects the explicit-

[3] Translation from Danish into English of a quotation in Weekend avisen, 9 November 1990, page 4 in WA Bøger.

ness, formal precision and accuracy typical of written language.

3.3 Standard Spanish and the language used in this corpus.

Standard Spanish has been defined as the language considered 'normal' and 'neutral' by the majority of the Spanish people, i.e. excluding "affected and vulgar expressions" according to Navarro Tomás[4].

Standard Spanish defined in this way is an ideal that can be approximated but can never be fully captured by any grammar.

According to this definition there will be several standards, for instance the Latin-American standard(s) and the Peninsular standard.

This text corpus, containing interviews from two magazines published in Madrid in 1990, may be said to illustrate the Peninsular standard in 1990.

3.4 Good usage and the language used in this corpus.

Closely related to the problem of standard Spanish is the problem of good usage, which needs neither be an ethical nor an aesthetic issue. In these texts, communication of ideas is an important purpose, and a classical definition of good usage is therefore useful: a text or sentence is written or spoken in good Spanish if, according to the purpose, the language is as concise as necessary and as simple as possible. From this point of view the present text corpus offers specimens of both bad and good Spanish.

The text corpus will be available free of charge for non-commercial research purposes from Kjær Jensen, The Aarhus School of Business, Fuglesangsallé 4, DK-8210 Århus V. (Phone +45 86 155588 / Fax + 45 86 157727).

---

[4] María Josefa Canellada / John Kuhlmann Madsen (1987), p.7.

Select bibliography:

Agencia EFE (1989): Manual de español urgente. Madrid. Ed. Cátedra.

Butt, John / Carmen Benjamin (1988): A New Reference Grammar of Modern Spanish. London. Ed. Edward Arnold.

Canellada, María Josefa / John Kuhlmann Madsen (1987): Pronunciación del español. Madrid. Ed.Castalia

Landau, Sidney (1989): Dictionaries. The art and craft of lexicography. Cambridge University Press.

Lauridsen, Karen M. / Ole Lauridsen (1989): "Tekstkorpora. En ny forskningsaktivitet ved Handelshøjskolen". In: Festskrift i anledning af Handelshøjskolens 50-års jubilæum 31. august 1989. Handelshøjskolen i Århus.

Lauridsen, Ole / Theis Riiber / Henning Søndergaard (1991): Erstellung eines dänischen und eines deutschen Textkorpus — Fachsprache Gentechnik. In: HERMES Nr. 6, 1991, pag. 125-138.

Martínez de Sousa, José (1985): Diccionario de ortografía. Madrid. Ed.Cátedra.

Real Academia Española (1931): Gramática de la lengua española.

Real Academia Española (1952): Nuevas normas de prosodia y ortografía.

Real Academia Española (1973): Esbozo de una nueva gramática española. Madrid.

Quirk, Randolph / S.Greenbaum / G.Leech / J.Svartvik (1985): A comprehensive Grammar of the English Language. Longman.