Measurements of Spoken Language Variability in a Multilingual Corpus. Predictable Aspects

Massimo Moneglia

LABLITA, University of Florence
piazza Savonarola 1, 50123 Florence, Italy
moneglia@unifi.it

Abstract

The paper provides cross-linguistic measurements of everyday language use based on the C-ORAL-ROM multilingual corpus of spontaneous speech. The average and the variation coefficient of a series of standard parameters are provided, faced to the main sociological and structural contexts of spoken language use. Mid-Length of Utterances (MLU); Mid-Length of the dialogic turn (MLTw); Speed; Mid length of the tone unit (MLTone); Fragmentation. Such variation parameters show strong predictable characters at cross-linguistic level. MLU has a positive correlation with MLTw and is shows highly predictable values in informal dialogic structures. Both MLU and MLTw have an inverse correlation with Speed. MLTone and Speed are predictable according to language specific features, but while MLTone have low intra-linguistic variation, Speed record a cross-linguistic tendency to lower values in formal language uses. Fragmentation is a permanent feature of spoken language, but it varies mainly according with speakers.

1. Introduction

One of the main characters of spoken language when compared to written language is the huge variability of the speech events according to individual characters, context of use, semantic domain. The paper provides a set of quantitative measurements of the range of natural variations in spoken romance languages.

The C-ORAL-ROM multilingual corpus of spontaneous speech (IST 2000-26228) offers a representation of the main contexts of use of the spoken domain for French, Italian, Portuguese and Spanish. In parallel, corpora are annotated with a set of relevant linguistic information (Cresti et al., 2002).

Each Romance corpus represents the main parameters used for the description of spoken language; that is: *Dialogue structure*, *Sociological domain of use*, *Genre*, *Semantic domain of application*; *Channel* (Labov, 1966; Biber, 1998; Biber, 1988; Gadet, 1996) with the same number of words, ensuring significance and comparability of the four collections (300,000 words each)

Language type	***Sociological context***	***Structure of the communication event***
Informal	Family/private Public	Dialogue Multi-dialogue Monologue

Language type	***Channel***	***Typical domain of use***
Formal	**Natural context**	political speech, political debate; preaching; teaching; professional explanation; conference; business; law
Formal	**Media**	talk shows; scientific press; reportage; interviews; sport; news meteo
Informal	**Telephone**	Private conversations; human-machine interactions

Corpora are transcribed and tagged with respect to:

- Dialogue structure; *Dialogic turns*
- Prosodic breaks; *Terminal, non terminal breaks and fragmentation events*
- Significant speech events: *utterances in the speech flow* (roughly 55,000 for each corpus)

Once the relevant information regarding the spoken events is marked in the speech resources, corpus based contrastive studies allow to foresee a significant series of correlations between the context in which a given spoken event occurs and its main linguistic qualities

2. Measurements

Cross-linguistic studies of standard measurements in the domain of Romance Languages have two main consequences. On one side tendencies that are consistent at a cross-linguistic level testify their linguistic significance; on the other side if the range of variation among romance languages turns out well defined then language specific characters can emerge.

Spoken language variability can be investigated by measuring the average and the variation coefficient of the following standard variation parameters (Cresti, 2000):

- Mid-Length of Utterances (in words)
- Mid-Length of the dialogic turn (in words)
- Speed (words per second)
- Mid length of the tone unit (in words)
- Fragmentation

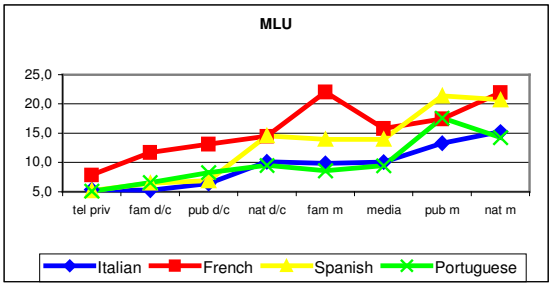
Language-dependent and language-independent values of such speech variation parameters can be highlighted through detailed histograms marking the correlations with the fields represented in the corpus design structure.

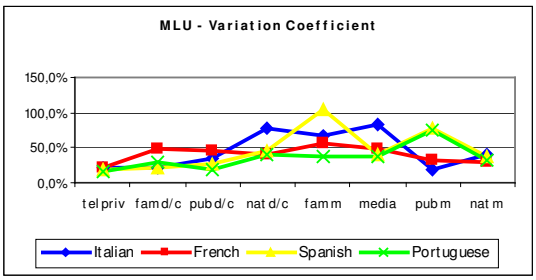
2.1. Mid-Length of Utterances

MLU has strong cross-language correlations with respect to all the variation parameters represented in the corpus structure:

- much higher in formal language
- higher in public context
- significantly higher in monologues
- variable in accordance with the channel (lower in media, with respect to formal in natural context, and in telephone, with respect to informal dialogues)

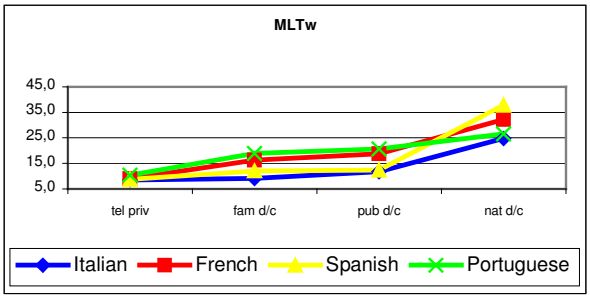
In informal dialogic structures MLU is highly predictable in all languages: Spanish, Italian and Portuguese vary around the same average (5-7) while French records a higher average (9-11). In the four collections the *variation coefficient* for MLU is constantly low in telephone (>20%) and in informal dialogues (from 20% to 25 %). On the contrary MLU is more variable in media and monologues. The variation coefficient is high in media emissions, in accordance with the emission format (over 50%). MLU may also vary consistently from text to text in both formal and informal monologues (variation coefficient around 40 %), probably in accordance with different strategies adopted for text organization.





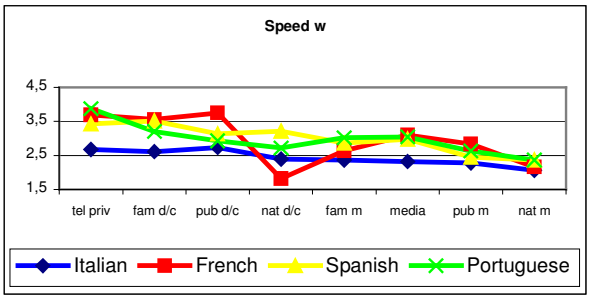
2.2. Mid length of the dialogic turn (MLTw)

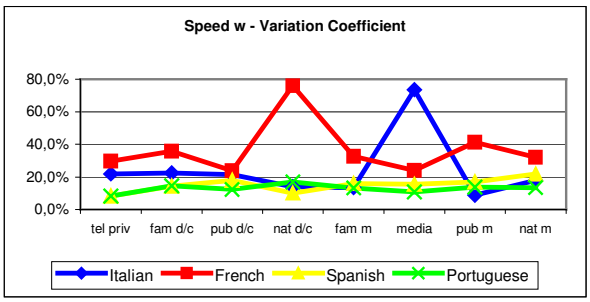
The MLU has a positive correlation with MLTw. Monologues, which have only one long turn, always have a higher MLU value, while in dialogue structures (telephone, family private, public informal and formal) MLU and MLTw co-vary in the four collections. Therefore in spontaneous speech the longer the turns of a text are the longer each utterance is.



2.3. Speed

Speed turns out to have an inverse correlation with MLTw and with MLU. A constantly higher value in telephone conversations and informal dialogues and a lower speed in the formal contexts can be observed cross-linguistically. The longer the turn the slower the flow of speech. Speed is a relevant predictable quality to discriminate the spoken styles of the four languages. In the informal dialogic sub-corpus (Telephone - Family – Public) each Romance language records a different speed: lower in Italian (2.7 w/s <), constantly over 3.5 w/s in French, between 3.5 and 3 in Portuguese and Spanish. The variation coefficient is reduced in telephone conversation and informal dialogues. On the contrary the average value of formal monologues discriminates the four languages to a lesser degree, still presenting a low variation coefficient (20% <). Therefore speed, which is limited by articulatory factors, is bound by language specific phonetic structures, however this difference is more evident in informal dialogues rather than in formal corpora, where speed is reduced in connection with the speech performance's task.





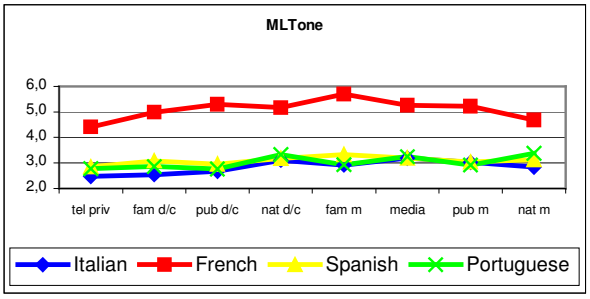
2.4. Length of the tone unit

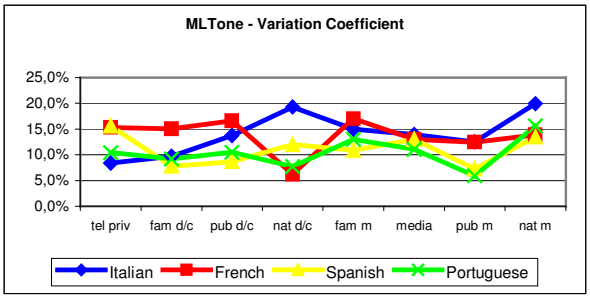
MLTone also shows strong upper limits (due to breath constraints) and, in theory, may vary in accordance with the prosodic, rhythmic properties of languages and their syllabic structure.

Contrary to MLU, MLTone cannot be predicted in accordance with the sociological and structural variation parameters.

The breath constraints force a low, predictable, average value. It is constant in Italian, Spanish and Portuguese (from 2.5 to 3.5), despite the fact that the former have a different prosodic structure (variation coefficient from 10 to 20 %). The strong variation which is found in French (from 4 to 6) is of course the main phenomenon to be explained.

The word weight in terms of number of syllables in French is probably at the origin of this variation. The syllabic reduction of words in speech, with respect to their graphic counterpart, is systematic in French. Therefore speakers can in principle produce more words within the breath unit.

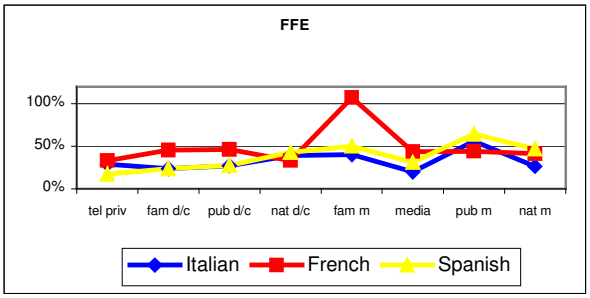




2.5. Fragmentation

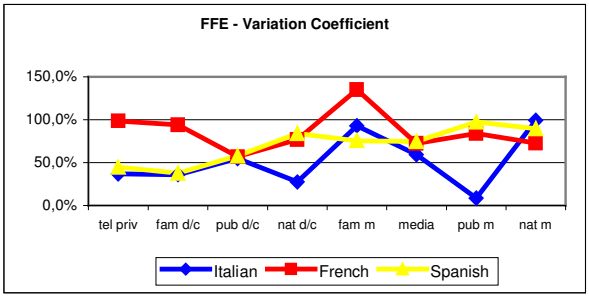
The incidence of fragmentation phenomena (interrupted utterances and false starts) when compared to the total utterance has been annotated for Italian, Spanish and French corpora, and is high in all collections. This testifies its relevance as a peculiar feature of spoken language. Fragmentation has an incidence that varies from 20% to 60% of utterances. Quite surprisingly, in all collections it has lower incidence (around 20%) in the informal dialogues, in telephone conversations and media emissions, while it ranges from 40% to 60% in the formal part.

Therefore, fragmentation turns out higher when the spoken performance is task oriented, while is relatively lower in everyday use. However although this parameter is predictable in average it is also strongly determined by individual factors as is shown by the lower incidence of fragmentation in media emissions (where professional speakers are involved) and confirmed by a high variation coefficient, all through the corpus nodes (from 40% to 100%).



3. Conclusions

In spoken language the utterance limits specify the domain of the main linguistic relations. For instance, argument structure, constituency, head dependency, and chunking relations, hold among elements of a same utterance.



The prevision of the length of utterance allows HLTs to consider the probability of occurrence of the utterance limits and therefore to better figure out the domain of actual linguistic relations. This may be significant for many purposes: adequacy of speech synthesis performances, selection of relevant domain for speech recognition and automatic indexing technologies.

The spoken language domain shows a strong variability with respect to this parameter, but certain variation tendencies can be clearly foreseen. The data presented show that in all the languages under investigation the Mid-length of the utterance varies according to the structure of the communicative event, the sociological domain of use, and the channel; this means that the length of the linguistic object which is the outcome of the speech act depends on non linguistic factors.

Very roughly speaking, non linguistic conditions specify the goals of the linguistic performance. Accordingly, the linguistic means are by preference limited to a few words or need complex structures with many words. More specifically the range of variation turns out quite predictable in informal dialogues, that is the prototypical domain of application of spontaneous speech. In informal dialogues the values are cross-linguistically recorded within two ranges (5-7 words per utterance in Italian, Spanish and Portuguese, around 10 in French) with a relevant variation between French and the other romance languages, specifically in the informal dialogic part. For all languages such values strongly diverge form the formal-monologic part, as they always record higher values. The Variation Coefficient with respect to MLU in the informal part is also much lower and testifies the significance of the correlation. Therefore the severe restriction of the number of words that belong to the same utterance in natural dialogic contexts must be seriously considered when dealing with speech.

The tendency to have much longer utterances is also cross-linguistically verified in connection to two features: a) monologic structure of the linguistic event; b) context requiring a formal use of language. However if the length of the utterance is predictable at higher values in those domains (always over 10 words per utterance), the range of variation is much higher. The variation is high specifically in media speech. No language specific tendency can be highlighted on the basis of the data at our disposal. In other words in order to foresee a narrow range of MLU in such non-prototypical domains the corpus sampling must deal with more subtle distinctions as genre and semantic domain. The C-ORAL-ROM corpus does not account for a sufficient amount of samples to document those variation .

Very significantly MLU is correlated with the length of the dialogic turn but, at least in general, not with the

Length of the tone unit. In all the languages in object the length of the Utterance and the length of the turn co-vary in accordance with socio-structural parameters. The more the context is formal the more the linguistic task requires a long turn and the more each utterance turns out long and structured.

This tendency seems "natural", but it must be pointed out that this is not in principle necessary. The reverse possibility could also be considered; that is, the more the linguistic performance is complex, the more each piece of information might be simpler. Apparently this theoretical alternative does not apply to natural languages.

Turn alternation in a given spoken dialogue is therefore a cue to predict the utterance length. On the contrary, in all the languages under consideration, the length of the tone unit appears independent from the contexts of use, and is fairly constant in all languages.

The variation in values of MLTone is relevant at cross-linguistic level. For what concerns Romance languages the similarity between Italian, Spanish and Portuguese turns out evident, while French strongly diverges, probably due to a different syllabic weight of words. The peculiar values of French seem therefore linked to internal structural properties of the language. The relation between phonetic syllables and tone units seems to be the basis of the relation between words and tone units.

The relation between MLTone and MLU must be studied carefully. While MLTone is constant throughout the corpus, MLU undergoes a strong variation. The two measurements do not co-vary, however it must be pointed out that French which has a higher MLTone has also a much higher MLU. Therefore, although the variation in MLU is not a function of MLTone, a higher MLTone may convey a higher baseline for MLU variations.[1]

The lack of a strict correlation between the length of the utterance and the length of the tone unit must be highlighted, for its linguistic significance. While the latter is a function of number of syllable and breath constraints the former it is independent from both factors. In other words the functional values of the utterance that vary in accordance with the tasks requested by the context of use are not strictly bound by the phonological structure of a language. In parallel, the tone unit does not convey a functional value related to the context.[2]

The results regarding Speed are quite interesting when compared with MLU and MLTone. Speed varies in accordance with both the corpus design and language-specific factors. Within the Romance family Speed turns out to be a character proper of each language, however this difference is mainly recorded in the informal dialogic part, where French and Spanish have clearly higher values. This is confirmed as a genuine language specific datum by the low Variation coefficient.

Cross-linguistic differences are lower in the formal part, where the speed decreases in all languages for reasons presumably linked to the task of the speech performance. Given this general tendency the variation in speed between the informal part and the formal part is more sensible for those languages with a higher speed.

However, speed variations are only partially correlated with variation in MLTone. French has indeed a higher MLTone and also a higher speed, however the two values are not correlated in Spanish, where MLU is lower but speed is high. Therefore other factors apart from the syllabic structure must determine such variation among the romance languages.

Speed has an inverse correlation with MLTw and with MLU: the longer the turn the slower the flow of speech.

The consistent frequency of fragmentation events with respect to the number of utterances in spoken language is confirmed by the available data. This is extremely relevant for speech analysis at both syntactic and prosodic level. All fragmentation events cause the onset of a PoS sequence that, by definition, is inconsistent with the grammar. Moreover all fragmentation events cause a prosodic rupture of the speech fluency and the onset of prosodic patterns that are also inconsistent with the prosodic rules. The frequency of fragmentation events determines sequences that have little or no meaning from an informational point of view. Given their percentile incidence their recognition is therefore vital for the understanding of spontaneous speech.

For what concerns measurements a significant similarity of incidence at cross linguistic level can be verified; that is fragmentation is a permanent datum of spoken language. However, despite this meaningful result, no significant correlation can be point out with respect to the corpus structure. Cross linguistically, we verified that non professional speakers have the tendency to give rise to more fragmentation in task oriented speech performance, rather in everyday language. However the high variation coefficient show that the tendency to fragmentation is more a speaker's character whose probability of occurrence cannot reasonably foreseen on the basis of the data considered.

[1] The correlation between MLU and MLTone in French is hard to be explained. Given that MLTone is constant, on the basis of the proposed hypothesis, the higher values of MLU in the Formal-monologic part of the variation should be proportionally higher with respect to the informal dialogic part. However this is not the case. Therefore supplementary data seems to be needed.

[2] This does not means that tone units does not convey any informational value. The variation in length of the tone unit may be related to the role played within the utterance.

4. References

Biber, D. (1988). Variation across speech and writing, Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Finegan. E. (eds.) (1998). Corpus linguistics: investigating language structure and use. Cambridge: CUP.

Cresti, E., 2000. Corpus di italiano parlato, vol. I- II, CD-Rom, Firenze: Accademia della Crusca.Cresti, E., Moneglia M., Bacelar, F., Sandoval, A. M., Veronis, J., Martin, Ph., Choukri, K., Mapelli, V., Falavigna, D., Cid, A., (2002). The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus. In Proceedings of LREC 2002 (pp. 2--10). Paris: ELRA.

Gadet, F., 1996. Variabilité, variation, varieté: le Français d'Europe. French Language Studies, 6, 45--58.

Labov, W., 1966. The social stratification of English in New York City. Washington D.C.