Speech and corpora:

How spontaneous speech analysis changed our point of view

on some linguistic facts

Philippe Martin

CLILLAC-ARP, EA 3967, UFR Linguistique
Université Paris Diderot Sorbonne Paris Cité
philippe.martin@linguist.jussieu.fr

Some 40 to 30 years ago, in the heroic days of corpus based research, transcription and acoustic analysis of either read or spontaneous speech was so time consuming that even the most ambitious projects would not exceed one or two minutes of recording. Furthermore, the linguistic analysis of this type of data was extremely rare, not only for technological reasons but for ideological reasons as well. Indeed, it was then considered that spontaneous speech was full of speaker "errors", disfluencies, logical gaps, and not worthy of scientific analysis.

In light of this, Claire Blanche-Benveniste and her GARS team in Aix-en-Provence made one of the greatest accomplishments by initiating and developing a project that would completely change our point of view on spontaneous speech and on many linguistic facts. Despite the technical and ideological difficulties encountered at the time, the GARS team gathered and transcribed a large set of spontaneous data, which is still being used and analyzed today. Slowly but surely, the published linguistic analysis pertaining to *la langue parlée* acquired a specific status among linguists, by proposing a macrosyntactic view of the speaker's spontaneous production and by demonstrating, with real life data, how some previously solid rock concepts were the result of misconceptions excluding the oral aspects of linguistic communication.

Thanks to this pioneer work, not only the usage of obvious spontaneous speech features such as hesitations, abandons, repetitions, reformulations, use of ponctuants, etc. became seriously documented, but the role of intonation could finally be established in a clear manner as an essential mechanism which ensures the final structuration of (micro) syntactically well-formed sequences of words in the sentence. This work is still going on, and many projects to elaborate spontaneous speech corpora of some considerable size have recently appeared in France and elsewhere. Among those developed since the GARS pioneer work, we can mention C-Oral-Rom, a project that assembled some 10 years ago similar speech data in four romance languages, and more recently, the C-Oral-Brasil project which carries on in the same spirit.

Among the many linguistic revelations brought by the analysis of spontaneous speech, we will mention and give some examples pertaining to French data. Conjunctions such as *parce que, puisque, etc.* constitute an interesting case where inherent properties of coordination or

subordination of these conjunctions can be overturned by intonation. Another example is given by the non-equivalence between final rise and fall in French interrogative sentences morpho-syntactically marked. The speaker's possibility to use a "prosodic correction" by adding syntactic segments not previously planned is another important property of spontaneous speech, which incidentally shows that prosodic and macrosyntactic structures are not necessarily congruent in the sentence.

Another essential point pertains to the emergence of the listener's point of view in the description of sentence intonation, leading to consider that listening to speech does not involve decoding of a sequence of words, but rather of sequences of syllables, in which prominent (stressed) syllables play an important role in signaling to the speaker the need to convert the last syllabic sequences into some higher rank linguistic unit. These chunks of syllables (usually called stress groups) constitute the basic elements of the sentence, which are merged at a later stage into larger units, possibly corresponding to syntagms, but not necessarily.

All these examples can lead us to reconsider the traditional view whereby syntax is the dominant and universal organizer of phonological or morphological units. Therefore, the ongoing emergence of projects assembling spontaneous speech data appears essential, as it may bring back the speaker and the listener(s) at the center of linguistic description.