Why we do not speak in Marseilles as they write in Paris? and consequently why NLP needs crucially more data from spontaneous speech ecological recordings

Professor Henri-José Deulofeu
Aix-Marseille University
Marseilles, France

Starting from the comparison of the structure of the same semantic content phrased in spoken ordinary French in (1a) and (2a), and in standard written French in (1b) and (2b),

1.a Moi ma mère tu sais le salon eh ben c'est de la moquette

Me my mother ye know the living room well it is carpet

1.b le sol du salon de ma mère est en moquette

the floor of my mother's living room is carpeted

2. a Alors cette soirée avec ton nouveau mec ? Eh bien Paul m'a emmené dans sa super voiture parce qu'il s'appelle Paul et nous sommes allés au cinéma voilà

So this evening with your new boy-friend? Well Paul took me with his super car because his name is Paul and we went to the movies that's all

2.b Mon nouveau compagnon s'appelle Paul. Il m'a emmené dans sa belle voiture et nous sommes allés au cinéma)

My new boy-friend is called Paul. He took me in his beautiful car and we went to the movies

I will rephrase the question in the title into two subquestions : (Subq1) why don't we speak as we write ? (Subq2) why we do not speak in Marseilles as they speak in Paris? I will first show that (Subq1) is not properly formulated. The source of the differences between variants (a) and (b) is not the generally admitted opposition between oral and written style, but indeed between spontaneous versus elaborated use of language. Basically, the function of spontaneous utterances (a) is primarily to establish a cooperative and convivial relationship between the participants, and only secondarily to convey an objective description of the situation. As a consequence, their overall structure is organized by prosodic dependency relations between grammatical "chunks". This formal "loose" structure is a suitable basis for a strategy of interpretation by means of pragmatic inferences, as it is the case for the pieces of a discourse (Deulofeu [2016]). By contrast, the main function of elaborated sentences (b) is to provide a picture of instances of real or possible worlds. The corresponding structures rely basically on grammatical dependency, which entails an interpretation strategy by means of semantic compositionality. Bringing additional evidence from ancient French, I will point out that there existed examples of spontaneous written French. This type of "discursive tradition" in the words of Kabatec (2011) has been progressively banned from written French along the history of French by professionals of writing (teachers, lawyers, journalists), who favored the representational function of language over the communicative one. As a consequence, the only available data for the study of French have been over a long period of time examples of elaborated French, ignoring the spontaneous varieties. French for grammatical studies was equated to a subset of possible utterances.

The problem for NPL is that, even with an increasing use of written spontaneous language through social media, we are far from having a sufficient data base to explore at deep the specific organization of spontaneous oral speech.

This lack of empirical basis will appear still more evident if we try to address Subq2 rephrased as : what are the structural differences between Northern and Southern varieties of French, if any? Letting obviously phonetics aside to focus on the basic grammatical structures of French, are we equipped enough to describe the different variants that are used everywhere in France? I will show first, using

data from typologist approaches[1], that it is possible to find in spontaneous spoken French as a whole all the structures of headed relative clauses present in other languages. And even some structures that have up to now been ignored in the literature. I will comment on one specific type, found in our Orfeo corpus spoken section :

(3) oui ça correspond au chiffre d' affaire qu' on faisait au début qu' il est arrivé hum chez nous

yes it corresponds to the turnover that we made at the beginning that he arrived hum in our business

I will then address the following issue : are some of these variants specific to southern varieties of French? I have no evidence of that for relative clauses. But if we extend our analysis to all possible non standard uses of the complementizer *que*, in a corpus driven study (Deulofeu 2013) , it appeared that *que* as a "universal" subordinator is more restricted in northern varieties than is southern ones. For instance, whereas we find in all varieties *que* introducing a subordinate clause conveying a contrast relation with the main clause :

(4) ils travaillent maintenant avec des machines que avant c'était tout fait à la main ça

They use now machines that (whereas) before it was all hand made, that (stuff)

our data doesn't show examples in Northern French of th following southern use of *que* introducing a consequence subordinate :

(5) ça nous fait des frais que là aussi les sous ils sont encore partis

Il entails expenses that (so that) the money it flew off again

But interestingly enough, I recently watch on a youtube tutorial a speaker with a clear Northern accent saying :

(6) Mettez les informations qui sont les plus importantes et un visuel qui soit attractif que si on arrive sur la page au bout de 3 secondes on sa- on sache de quoi il s'agit

Just put the most relevant informations and a visual that is attractive that (so that) if someone arrives on the page after 3 seconds he can know what is going on

Contrary to the data available before, the *que* in (6) introduces clearly a consecutive subordinate.
If we put together the intermediate conclusion of the discussion of Subq 1 and the conclusion we can draw for Subqu2, we can argue for the general conclusion that we lack the relevant data both for completing the linguistic description of French and for training the tools of NLP. This urge us to continue accumulating data from spontaneous language transcripts if we want that the corpora used in NLP be a relevant picture of all the possible variants of the French language, all the more that we have at last the tools to make it easy and financially asequible.

REFERENCES

DEBAISIEUX J.-M. (eds) (2013). *Analyses linguistiques sur corpus : subordination et insubordination en français*. Paris : Hermès / Lavoisier.

DEULOFEU J. (2013). « *que* dans les phénomènes de subordination », in DEBAISIEUX (2013), 365-408

DEULOFEU J. (2016). La macrosyntaxe comme moyen de tracer la limite entre organisation grammaticale et organisation du discours. *Modèles linguistiques* 74, 135-166.

KABATEC J., (2011) : « Diskurstraditionen und Genres » in Sarah Dessì Schmid, UlrichDetges, Paul Gévaudan & Wiltrud Mihatsch (éd.), Rahmen des Sprechens :Beiträge zu Valenztheorie, Varietätenlinguistik, Kreolistik, kognitiver und historischer Semantik ; Peter Koch zum 60. Geburtstag, Tübingen

ORFEO : https://orfeo.ortolang.fr/?locale=fr

[1] World Atlas of Language Structures : https://wals.info/