Age, sex and first language for the forensic linguistic profiling of teenagers in Catalan

Roser Giménez García

Margarita Salas Postdoctoral Research Fellow
University of Barcelona
Barcelona
Spain

Awarding Institution: University of Barcelona, Spain
Date of Award: 26 May 2022

KEYWORDS: sociolinguistics, forensic linguistics, linguistic profiling, age, sex, first language, Catalan
Research in forensic linguistics on minors’ use of language has often focused on discourse in investigative interviews (e.g. Jol and Stommel 2021). For instance, according to the Spanish National Statistics Institute, 11,238 minors between 14 and 17 years of age were convicted by criminal courts and tribunals in 2020 (INE 2021), and these minors would ordinarily have participated in investigative interviews in the lead up to their trials. Moreover, children and adolescents come into contact with law enforcement agencies and the justice system not only when they are charged with crimes but may also become involved in the fact-finding process at other stages, and their language use is therefore not confined to investigative interviews. For example, in online grooming or bullying cases, samples of the language used by minor victims or witnesses may play a central role as evidence throughout the investigation of the facts and eventually in court. Therefore, forensic linguists may be asked to assist in the investigative or judicial phases by performing a range of tasks on samples of the language of minor speakers, including authorship analysis or linguistic profiling. It should be noted, however, that the preparation of informative and reliable linguistic profiles relies on pre-existing descriptive studies of language varieties in use by different social groups, e.g. adolescent speakers (cf. Nini 2014: 16); such studies are limited, particularly with regard to Catalan-speaking children and adolescents.

Catalan sociolinguistics has paid some attention to the spoken informal language of young speakers. Nonetheless, these studies have often focused on the impact that the language contact between Spanish and Catalan has on these varieties, especially regarding phonetic and lexical phenomena (e.g. Ballart 2004; Bernal and Sinner 2009). Scholars have also often analysed data from radio or television programmes and only rarely from unplanned exchanges (e.g. Wieland 2009). In addition, the application of research on Catalan to forensic purposes is still in its infancy, with very few studies ploughing this field so far (Forensiclab 2013). Indeed, numerous lines of research are yet to be explored, including the empirical substantiation of the forensic linguistic profiling of questioned samples in Catalan. This panel study aims to fill this gap in the literature by exploring the relationships between linguistic and extralinguistic variables in the language of adolescent speakers in order to inform forensic linguistic profiling practices.

To do so, this panel study analysed a sample of the speech of 69 adolescent speakers from two Catalan cities through the lens of variationist sociolinguistics and examined the research question ‘are there linguistic variables that may be used to identify the most likely age, sex and first language in adolescent speech in Catalan?’ The data used were taken from a pre-existing longitudinal research project on the language competence, practices and attitudes of over 2500 students across the Catalan-speaking areas in Spain (Bretxa et al. 2016). Of those, 120 were selected to carry out two tasks designed to test their spoken skills. The first task
was intended to reproduce an academic setting, whereas the second task tested students’ interactive skills through a role-play activity in which they embodied the salesclerk in a pet store and their communicative goal was to persuade the interviewer-customer to make a purchase. The sample used in this study comprised the recordings and transcripts of 69 students for whom data were available. My colleagues in the Center for Research in Sociolinguistics and Communication of the University of Barcelona (CUSC-UB) had recorded them when they were 12 years old and then again at 16. Only the second task was used, since the register in these samples is more directly comparable to the informal registers usually found in forensic data. A total of 21 linguistic variables were analysed in relation to speakers’ age (12 or 16 years of age), sex (male or female) and first language (Catalan, Spanish or both) using descriptive statistical measures (the minimum and the maximum, the mean, the median and the lower and upper bound of the confidence interval for the median), Principal Component Analysis (PCA), tests of statistical significance (the Wilcoxon signed-rank test, the Mann–Whitney U test and the Kruskal–Wallis test by ranks; Scheff 2016; Van Buren and Herring 2020) and classification trees.

The linguistic variables analysed in this study were selected from the available literature on variation phenomena in Catalan and Spanish (e.g. Vila i Moreno 1998; Illamola and Queralt 2020) and studies on language acquisition in adolescence (e.g. Nippold 2004). They were divided into pragmatic, morphosyntactic and lexical variables. The pragmatic variables include eight pairs of discourse particles formed by lexical loans from Spanish and their corresponding Catalan counterpart (a ver/a veure ‘let’s see’, bueno/bé ‘well’, después/després ‘then’, hombre/home ‘man’, o sea/o sigui ‘that is’, pero/però ‘but’, pues/doncs ‘so’ and vale/d’acord ‘okay’), greetings (hola ‘hi’, bon dia ‘good morning’, bona tarda ‘good afternoon’, hola bon dia ‘hi good morning’) and instances in which the speakers reproduce elements first produced by the interviewer (heterorepetitions). The morphosyntactic variables are the location of clitic pronouns in periphrastic constructions (before, after or before and after the verb phrase), the number of non-rhetorical questions, question types (open- or closed-ended), conditional constructions (introduced by si ‘if’, mentre(s) ‘as long as’, a no ser que ‘unless’ or en (tot) cas ‘in any case’), obligation periphrases (tenir que + infinitive, haver de + infinitive, tenir de + infinitive or caldre + infinitive2), demonstrative determiners3 and plural dative clitic pronouns (els hi, els, li or lis4). Lastly, lexical variables include contact language phenomena (importations, i.e. loanwords and code-switching, hybrid elements, calques and others), colours (blau ‘blue’ or de color blau ‘blue-coloured’) and the number of adverbs ending in -ment (‘-ly’).

PCA is a multivariate analysis technique which transforms large sets of correlated variables into more manageable, two-dimensional models. In this study,
PCA accounts for percentages between 29.8 and 56.3 of the variance for all the linguistic variables and identifies 16 potentially significant relationships between the dependent and specific linguistic variants. The results of the tests of statistical significance show that 14 linguistic variables significantly correlate with speakers’ age, ten with first language and six with sex. Using the linguistic variables in the study, classification trees reach a correct classification rate of 85.5% for speakers’ age, 78.3% for sex and 70.3% for first language.

Thus, the research question is answered affirmatively, since the combination of statistical techniques confirms that the linguistic variables analysed here can be used to successfully identify the most likely age, sex and first language of adolescent Catalan users. This pioneering exploration of the use of language by adolescent speakers for forensic linguistic purposes opens up many unexplored avenues for further interdisciplinary research in Catalan sociolinguistics, such as the analysis of online data (e.g. for the detection of the impersonation of a fake adolescent identity by an adult), the study of other independent variables (e.g. ethnic origin or geographical area, Queralt 2014: 37–38) or the application of mixed methodologies to the description of adolescent language use.

Acknowledgements
I am deeply indebted to Dr Sheila Queralt, Dr F. Xavier Vila and Laura Barrios for their help and encouragement, and to Dr Núria Alturo, Dr Miquel Àngel Pradilla and Dr Jordi Cicres for their insightful comments. This work was supported by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the European Social Fund (2021FI_B2 00016).

About the author
Roser Giménez García is Margarita Salas Postdoctoral Research Fellow, University of Barcelona. Contact address: Laboratorio SQ-Lingüistas Forenses, Pallars, 193, Barcelona 08005 Spain. rosergimenez@ub.edu

Notes
1. Following authors like Serrano (2008: 177), sex is conceptualized as a biological binary variable based on genetic, organic and hormonal characteristics used to differentiate between males and females, whereas gender is understood as a socioculturally acquired dimension related to behavioural and identity patterns upon which the social distinction between men and women is established (cf. Breu 2013a, 2013b). The main data collection tool used in the pre-existing longitudinal research project consisted of a comprehensive sociolinguistic survey, which asked students about their sex but not their gender identity. Thus, only sex is considered in this study.
2. All of these constructions are approximately equivalent to have to + infinitive.
3. The variants of this variable include uses of *aquest* ‘this’ and *aquell* ‘that’ as adjectives, appearing before and after nouns (e.g. *aquest animal* and *l'animal aquest* ‘this animal’), and as pronouns (e.g. *aquest* ‘this one’ or *un animal d'aquells* ‘one of those animals’).
4. All these pronouns translate to ‘to them’. However, only *els* is described as acceptable in all registers of the language in the prescriptive grammar. *Els* *hi* is considered acceptable in informal registers. *Li* and *lis* are notably marked as originating from the language contact situation with Spanish and are not considered as acceptable forms in prescriptive works. All forms, especially the latter three, are found in informal registers. This variable is explored in detail in Giménez García (2023).

References


