Person effects in null / pronominal subject alternation in Romanian: a corpus study

Fabian Istrate, Anne Abeillé, Barbara Hemforth
LLF, Université Paris Cité

So-called pro-drop languages vary considerably with respect to the frequency of null subjects. Moreover, the alternation between pronominal and null subjects seems to be sensitive to various factors in Romance languages (Mayol, 2012). This paper adds Romanian data from corpus studies to shed more light on these factors. In a corpus study on spoken Brazilian Portuguese, a language having shown a decrease in the use of null subjects over time (Duarte 2000), Correa Soares et al. (2020) found that pronominal subjects were more frequent with 1st and 2d person (72%) compared to 3rd person (39%). While null subjects are generally more frequent in European Portuguese, Duarte (2000) also shows a higher frequency of pronominal subjects for discourse persons (35% for 1st, 24% for 2nd pers.) compared to 3rd pers. (21%) in a spoken corpus. Similarly, in spoken Spanish, Manjón Cabeza-Cruz et al. (2016) and Ávila & Segura Lores (2022) found a higher frequency of 1st pers. sing. pronominal subjects (Granada: 24.7%, Malaga: 33.5%) compared to 3rd pers. sing. (Granada: 10.6%, Malaga: 7.69%).

We annotated extracts from a written Romanian corpus (Parseme-ro 1.2, from the Agenda newspaper: 447,464 sentences, 13M. words) and a spoken corpus (CoRoLa: 152 radio recordings) (Barbu Mititelu et al., 2018). The average sentence length is about 18 words in both corpora. We extracted 400 non-embedded sentences, 200 from text and 200 from speech, half with null, half with pronominal subjects (1, 2) and manually annotated them with: subject type, person, animacy, number, gender, verb lemma, voice, polarity and animacy.

The most interesting effects were those related to person, which differ considerably from previous corpus studies on other Romance languages: logistic regressions show a significant main effect of person (β = -2.85, sd = .30, z=-9.6, $p < .001$), with a much higher frequency of null subjects for discourse persons in particular in the written corpus (interaction: $\beta = 3.80$, sd = .59, z=6.4, $p < .001$). This tendency can (among others) be accounted for by Ariel's (1990) Accessibility Theory: the more salient a referent is, the less explicit the subject will be. Since dialogue persons are inherently human, they are more prominent in discourse. It is also in line with Dobrovie-Sorin & Giurgea (2013)'s suggestion that pronominal subjects are used in Romanian to mark contrast and emphasis, and to avoid gender ambiguities (only 3rd pers. pronouns are marked for gender). However, saliency cannot account for the inverse pattern found in other Romance languages. It is thus possible that different factors play a role across Romance languages. Larger and more fine-grained parallel corpus studies as well as controlled crosslinguistic experiments will be necessary to shed more light on these differences.

Gender, animacy, number, and agentivity did not play a statistically reliable role. In the general model, voice did not have a significant effect either. However, we are facing a sparse data problem here because of the limited occurrences of non-active voice in both corpora (12% non-active). We therefore looked at active and non-active voice cases separately (see Figures 2a,b). Pronominal subjects are more frequent than null subjects for non-active voice (β=-1.37, sd=.42, z=-3.23, $p<.01$). This preference is marginally stronger in the written corpus (β=1.43, sd=.80, z=1.78, $p<.08$). No significant differences in frequency of null and pronominal subjects were found for active voice. Because of the sparse data problem, more controlled experimental studies will be useful here as well.

Even though more data will be necessary, we conclude that Romanian subject alternation is sensitive to person (and possibly to voice). These results are generally more pronounced in written than in spoken corpora. We assume that the norm for the use of null subjects is playing a

role here in particular for Discourse Persons. Editing processes aiming at more explicit gender disambiguation may have increased the use of pronominal subjects for 3rd person in the written corpus. Ongoing corpus research (embedded clauses) will provide further evidence about factors responsible for subject alternation, which are necessary to explain the difference between Romanian and other pro-drop languages.

(1) Deci nu pentru bani am ales-o.
so not for money AUX.1SG chosen.PST.-CL.3SG.F.ACC
'So, I haven't chosen her for the money.' (CoRoLa, 2014)

(2) Ea va rămâne deschisă la Timișoara până în 15 mai.
she will remain open.SG.F at Timisoara until in 15 May
'She will still be opened, in Timisoara, until the 15th of May.' (Parseme-ro, 2020)

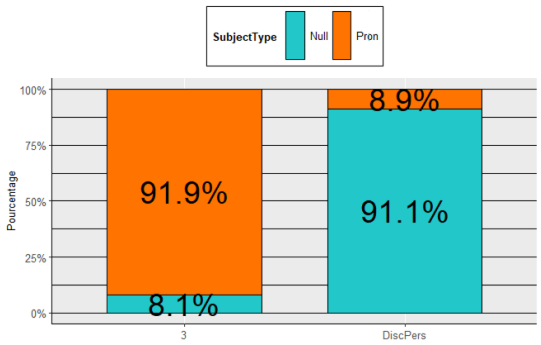


Fig 1a. Person effect in Romanian written corpus

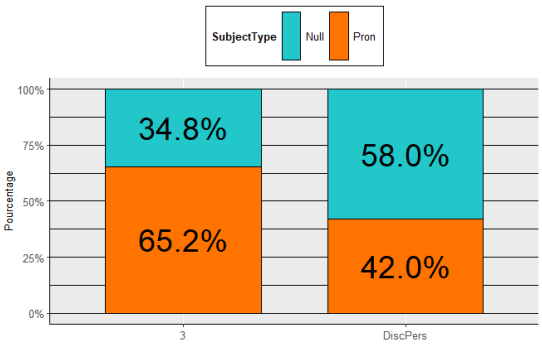
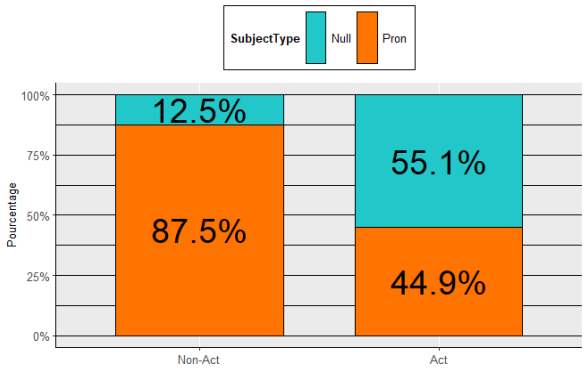


Fig 1b. Person effect in Romanian oral corpus



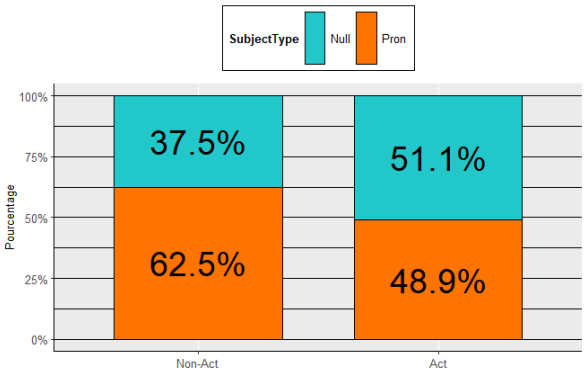


Fig 2a. Voice effect in Romanian written corpus Fig 2b. Voice effect in Romanian oral corpus

Selected References:

Ariel, M. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.

Ávila, A. M . & Segura Lores, A. 2022.Estudio de las variables predictoras de la expresión del sujeto pronominal en el corpus PRESEEA. Málaga. Nivel de instrucción bajo, *Anuario de Letras. Lingüística y Filología*, 10(2), 57-93.

Correa Soares, E., Miller, P. & B. Hemforth 2020. The Effect of Semantic and Discourse Features on the Use of Null and Overt Subjects - A Quantitative Study of 3rd Person Subjects in Brazilian Portuguese. *DELTA* 36(1).

Dobrovie-Sorin C., Giurgea I. (eds) 2013. A reference grammar of Romanian, vol.1, J. Benjamins.

Duarte, M. 2000. The loss of the Avoid Pronoun principle in Brazilian Portuguese. In Kato & Negrão (eds), 17-36.

Mayol, L.(2012). An account of the variation in the rates of overt subject pronouns in Romance. Spanish in Context, 9:3, 420–442.

Manjón-Cabeza Cruz, A., Pose Furest, F., & Sánchez García, F.,J., 2016. Factores determinantes en la expresión del sujeto pronominal en el corpus preseea de Granada. *Boletín de filología*, *51*(2), 181-207.