Chapter 6

Next mention biases predict the choice of null and pronominal subjects

Fabian Istrate[a], Ruxandra Ionescu[a] & Barbara Hemforth[b,a]
[a]Université Paris Cité [b]CNRS

In this paper we investigate the production of null and pronominal subjects in Romanian. Data from corpus and experimental studies show that several factors need to be taken into account for this alternation beyond the syntactic role of the antecedents. Subject pronouns in Romanian seem to be sensitive to semantic and pragmatic factors influencing accessibility or predictability, induced among others by the interaction between verb semantics and coherence relations. Our results contribute to the larger debate regarding the extent to which predictability affects the choice of referring expressions, suggesting that null subjects are preferred not only for a subject antecedent, but also for a referent that is more predictable in the context.

1 Introduction

1.1 Pronoun resolution and predictability

Pronoun resolution (i.e., the question of how pronouns retrieve their antecedents in discourse) is one of the most studied phenomena in psycholinguistics (see, for recent studies, Holler & Suckow 2016, Arnold 2010, Kehler & Rohde 2019, Schulz et al. 2021, Colonna et al. 2012). While much of the empirical work concerns processes in comprehension, a closely related question is which type of referential expression a speaker will use to refer back to a specific antecedent (Fukumura et al. 2022). In this paper, we aim to add both corpus and experimental evidence about the division of labor between null and pronominal subjects in a pro-drop language like Romanian. In particular, we are interested in investigating which

Fabian Istrate, Ruxandra Ionescu & Barbara Hemforth. 2025. Next mention biases predict the choice of null and pronominal subjects. In Gabriela Bîlbîie & Gerhard Schaden (eds.), *Empirical issues in syntax and semantics: Selected papers from CSSP 2023*, 139–170. Berlin: Language Science Press. DOI: 10.5281/zenodo.15450438

factors affect speakers' choices between these referential expressions and how next mention biases (i.e., the probability of mentioning a particular referent next) influence their production (Arnold 2001).

As a general principle of communication, when referring back to a previously mentioned and highly accessible entity in discourse, speakers tend to use shorter or less specific forms of reference, while making their exchanges as informative as necessary. This tendency to make the communication as economical as possible was captured by Grice (1975) in his *Maxim of Quantity*. Communication becomes more efficient as it requires less speaker effort without generating significant communicative cost (see also Levy & Jaeger 2007). To ensure that a referent is still accessible to the listener, the entity should be in the focus of attention, making it highly predictable to be mentioned next. A current debate in the literature addresses the question whether a more predictable referent will be recovered by a pronoun or a more complex referential expression (such as proper names, definite descriptions, etc.). In this respect, previous studies offer contradictory results: while some of them show no evidence that pronoun production is affected by referent predictability (Ferretti et al. 2009, Fukumura & Van Gompel 2010, Kehler & Rohde 2013, Rosa 2015, Patterson et al. 2022, among others), others suggest the existence of such an effect (Arnold 2001, Rosa & Arnold 2017, Lindemann et al. 2020, among others). Theories about the choice of a given referential form can be traced back to earlier hypotheses proposed in the literature on pronoun resolution. Givón (1983)'s *Topic Continuity Theory* posits that the degree of topicality that a given entity has in discourse is correlated with the usage of referential forms. In subsequent work, Ariel (1994) proposed in her *Accessiblity Theory* that referential forms can be ranked on an explicitness scale, ranging from full names to zero expressions. The more accessible an entity is in discourse, the less complex its corresponding anaphoric expression will be. In line with these hypotheses, von Heusinger & Schumacher (2019) put forward the concept of *prominence* aiming to characterize *a highlighted entity* in discourse. The properties of a prominent referent are based on the following criteria of prominence in grammar, given by Himmelmann & Primus (2015): (i) linguistic units of equal rank compete for the status of being in the center of attention, (ii) their status may shift, (iii) prominent units act as structural attractors in their domain. They suggest that the focus of attention can be shifted by linguistic means, thus updating the prominence structure. Similar observations postulate that semantic/coherence-driven predictability of referents in discourse may come from multiple sources, inducing next mention biases (see Gernsbacher & Hargreaves 1988, Gordon et al. 1993, Grosz et al. 1995, Fukumura & Van Gompel 2011): intrinsic properties like animacy, grammatical factors like subjecthood, information structural factors like topichood, or semantic and pragmatic factors like verb semantics or coherence relations.

1.2 Null and pronominal subject alternation in pro-drop languages

A prominent account of the alternation between null and pronominal subjects is the *Position of Antecedent Hypothesis* (henceforth, *PAH)* proposed by Carminati (2002) for intra-sentential anaphoric subjects in Italian. The *PAH* posits that null subject pronouns generally refer back to subject antecedents (1a) whereas pronominal subjects usually prefer a non-subject antecedent (1b).

(1) a. Marta scriveva frequentemente a Piera quando era negli
Marta write.pst.ipfv.3sg frequently to Piera when was in
Stati Uniti.
States United
'Marta wrote to Piera often when she was in the United States.'

b. Marta scriveva frequentemente a Piera quando **lei** era
Marta write.pst.ipfv.3sg frequently to Piera when she was
negli Stati Uniti.
in States United
'Marta wrote to Piera often when she was in the United States.'

A similar pattern has also been observed by de la Fuente & Hemforth (2013) for Spanish, showing the impact of grammatical role of the antecedents. However, in their experimental study, tendencies in Spanish seem to be less strong than those found for Italian, i.e. null subjects can also take object antecedents. Moreover, aside from the subject function of antecedents, de la Fuente & Hemforth (2013) point out that left-dislocation of the antecedents increases their discourse accessibility, thus being mostly retrieved by null subjects (see Runner & Ibarra 2016 for similar observations about information structure). In a production and a comprehension experiment on Greek and Italian, Torregrossa et al. (2020) confirmed the left-dislocation effect by arguing that null subjects have a bias for left-dislocated objects compared to *in-situ* objects. In line with the previous studies, Contemori & Di Domenico (2021) postulate that Italian and Spanish display a distinct division of labor between null and pronominal subjects: whereas in Italian the production of the referential form seems to be very distinct according to the grammatical role of the antecedent, in Spanish the division of labor is less clear. The results of their production experiment show that verb bias and causal coherence relations might play a role in Italian and in Mexican Spanish, although to a lesser extent in the latter: speakers produce a null subject to corefer to the object when an object-biased verb is present in the context. However, Chamorro (2018) suggests different tendencies for Spanish in an offline judgment task and an online eye-tracking study showing that null subjects do not exhibit a clear preference while pronominal subjects mostly prefer object antecedents. She also postulates that clause order (i.e., main-subordinate vs. subordinate-main) may also be responsible for the antecedent preferences of pronouns in the case of intra-sentential anaphora. Regarding the choice between a null and a pronominal subject in Romanian, Lindemann et al. (2020) and Istrate et al. (2022) have shown that grammatical role plays an important role in this alternation, the tendencies being close to those found for Spanish (de la Fuente & Hemforth 2013). Moreover, Lindemann et al. (2020) argues that null subjects are the most preferred choice when referring back to a more prominent or accessible referent, i.e. a subject antecedent which is also the *goal*, in terms of thematic roles. Divergent results concerning the production of these referential expressions might be due to different experimental tasks (online vs. offline), but also to existing differences in the experimental material. Further parallel corpus and experimental studies are necessary in order to make a proper crosslinguistic comparison between pro-drop languages.

1.3 The current study

The goal of the current study is to add evidence about the production of intra-sentential anaphoric subjects combining corpus and experimental data on Romanian, a pro-drop language in which the choice of null and pronominal subjects has not been extensively studied from a quantitative perspective so far. In particular, we are interested in investigating which role predictability and prominence may play in pronoun production, by shedding light on the potential influence of grammatical and semantic-pragmatic factors. In Section 1, we present some background about pronoun resolution and predictability, followed by empirical evidence on the choice of referential expressions in pro-drop languages. Then, we focus on Romanian as a consistent null subject language and we put forward our hypothesis regarding Romanian based on previous studies. In Section 2, we present a corpus study on complex sentences (sampling, annotation, results and discussion). Given the results we found in particular with respect to discourse relations, we ran a follow-up experiment in order to test to what extent the choice of referential expressions is sensitive to predictability invoked by implicit causality verbs in causal relations (Section 3). In Section 4, we put together the corpus and experimental results, by pointing out that preferences in the production of null and pronominal subjects in Romanian go beyond the grammatical role of the antecedents and must include predictability and discourse relations.

1.4 Hypotheses about Romanian as a pro-drop language

Romanian is a language that licenses the presence of null subjects (Dobrovie-Sorin & Giurgea 2013). In this subsection, we firstly present the classification of pro-drop languages proposed by Holmberg (2010), which underpinned the predictions made for Romanian. Consistent null subject languages (Italian, Spanish, European Portuguese[1], Greek, etc.) permit the use of a null subject irrespective of number, person or verb tense. A second category includes partial pro-drop languages (such as Russian), which limit null subjects to the 1st and 2nd person in finite clauses, and 3rd person pronouns *bound by a higher argument.* A third category consists of expletive null subject languages (such as German), allowing null expletive subjects but not referential ones.[2] The last category is represented by radical pro-drop languages (or *discourse pro-drop languages*, such as Japanese or Chinese), which permit other nominal arguments (e.g. objects) to be null, in addition to null subjects. According to Holmberg (2010), Romanian falls into the category of consistent null subject languages. In Romanian, there are several ways in which speakers may refer to an entity, including null subjects (2a), but also overt subjects realized as personal pronouns (2b), demonstratives (2c), proper names (2d), definite descriptions (2e), etc.[3]

(2) a. A ajuns la petrecere.
has arrived at party
'He/she has arrived at the party.'

b. *El* a ajuns la petrecere.
he has arrived at party
'He has arrived at the party.'

c. *Acesta* a ajuns la petrecere.
this has arrived at party
'This one has arrived at the party.'

[1] The pro-drop status of Brazilian Portuguese is controversial (see Duarte 1995, 2000 for *ongoing parameter change* of Brazilian Portuguese).

[2] German has more recently also been described as a topic drop language, see Schäfer (2021).

[3] In this paper, we will make the distinction between pronominal subjects (personal pronouns) and lexical subjects (proper names, definite descriptions) for methodological reasons regarding the annotation of the collected data.

d. *Alexandru* a ajuns la petrecere.
Alexandru has arrived at party
'Alexandru has arrived at the party.'

e. *Colegul nostru* a ajuns la petrecere.
colleague poss.1pl has arrived at party
'Our colleague has arrived at the party.'

With respect to consistent null subject languages, the alternation between null and pronominal subjects has attracted particular attention in the linguistic as well as the psycholinguistic literature (Carminati 2002, Chamorro 2018, Torregrossa et al. 2020, Lindemann et al. 2020, Contemori & Di Domenico 2021). According to Ariel's (1994) *Accessibility Theory*, null and personal pronouns (either stressed or unstressed) are very close on the accessibility scale, compared to definite descriptions or proper names. Alternation between these two referential expressions therefore needs quantitative research to shed light on the sometimes fine-grained distinctions determining their choice. Thus, we will only focus on these two referential expressions, in line with previous work.

Based on previous work on Romanian (see Lindemann et al. 2020), we hypothesize that null subjects will be preferred not only for subject antecedents, but also for more prominent or predictable referents in the context. We seek to establish: (i) in how far verb semantics and discourse relations render antecedents more predictable, and (ii) to what extent the referent predictability impacts on pronoun resolution in Romanian (see Demberg et al. 2023). From a comparative perspective, we expect to observe a similar pattern in Romanian as in Spanish, Italian (Contemori & Di Domenico 2021) and Catalan (Mayol 2018): null subjects should mostly be produced when the choice of the referent in the upcoming subordinate clause is in line with next mention biases. More precisely, we predict that the implicit bias of a verb for an upcoming referent makes the choice of a null subject for this referent more likely. However, if our predictions are on the right track, Romanian should be different from Mandarin Chinese, *a discourse pro-drop language*, where no evidence has been found for next mention biases affecting the production of referential expressions (Hwang et al. 2022).

2 Corpus study

2.1 Details about Romanian corpora

We used two corpora in our studies: the *Parseme-ro 1.2* corpus for written Romanian and the *CoRoLa* corpus for spoken Romanian. The *Parseme-ro 1.2* corpus

(Savary et al. 2018) is a written corpus of texts collected from the *Agenda* newspaper (containing 56,703 sentences and 1,015,624 words). Although the corpus has no subsections, it is a homogeneous journalistic corpus. Some texts included in *Parseme-ro 1.2* are also part of the *Romanian Universal Dependencies* corpus.

The *CoRoLa* corpus (*The Reference Corpus of the Contemporary Romanian Language*, Barbu Mititelu et al. 2018) comprises a written part and an oral part. In order to compare the production of null and pronominal subjects in Romanian, we extracted data from the oral part of the *CoRoLa* corpus (covering 151 hours, 57 minutes and 21 seconds). The oral texts in *CoRoLa* are mainly professional recordings from various sources (radio stations, recordings) for which transcriptions are available. Another part of the oral corpus is represented by texts read by various speakers in various circumstances: news read on radio stations, texts read by people close to them and texts read by professional speakers recorded in studios. However, we focused on spontaneous speech, taking into account only extracts from radio news and interviews. The reason for this choice was to have two sufficiently different sub-corpora so that we could expect to find interesting effects. We did not find any a priori reason to believe that read texts should be particularly different from written texts.

2.2 Corpus sampling

As the *Parseme-ro 1.2.* corpus is morpho-syntactically annotated, we used SQL query formulas to collect the data, both for null and pronominal subjects. However, for the *CoRoLa* corpus there is currently no such automatic annotation. We extracted occurrences with null and pronominal subjects, using the most common verbs in Romanian (a total of 552 verbs).[4] We thus constructed a sample of 368 complex sentences.[5] Following Oakhill et al. (1989), who point out the role of semantic and pragmatic effects of main clause factors in pronoun resolution, and in line with previous studies (Soares et al. 2020, Costa et al. 2004, de la Fuente & Hemforth 2013), we analyzed the production of null and pronominal subjects occurring in subordinate clauses. Moreover, since main sentences generally did not provide information about the previous antecedent in the context (such as syntactic function), we decided to analyze the choice of null and pronominal subjects in subordinate sentences. We excluded a number of occurrences in which the alternation between null and pronominal subjects was not possible in the context

[4] While lexical subjects certainly play a role for referential expressions, we only focused on null and pronominal occurrences in this study in line with much of the experimental work on this topic.

[5] The production of null and pronominal subjects in simple clauses is part of a separate study.

or in which annotation of our factors of interest in this study was not possible. More precisely, we did not retain in our study sentences that met the following criteria: (i) when a null subject was impossible in the context, i.e. pronominal subjects of non-finite verbal forms (infinitives and gerunds), as in (3a); (ii) when null subjects or pronominal subjects were discourse persons (1st or 2nd person) as in (3b), since it is often impossible to establish information about the factors of interest such as the grammatical role of the antecedent;[6] (iii) when null subjects were the only option in the context, i.e. null subjects of impersonal reflexive verbs (3c).

(3) a. Când ne întâlnim, vorbim pe ungurește, el fiind
when cl.1pl.acc meet.prs.1pl talkprs.1pl prep Hungarian he being
pe jumătate maghiar.
prep half Hungarian
'When we meet, we speak Hungarian, as he is half-Hungarian.' (corola-38914)

b. Dar când trebuie să fiu eu, atunci roșesc, ne
but when have sbjv be.sbjv.1sg I then blush.prs.1sg cl.1pl.dat
mărturisește.
confess.prs.3sg
'But when I have to be myself, then I blush, he confesses.'

c. Deși starea carosabilului este deplorabilă, nu se pot
although state.def road.def.gen is deplorable neg refl.3 can
face lucrările de reabilitare necesare, din lipsa fondurilor.
do work.pl.def of repair necessary of lack funds.def.gen
'Although the state of the road is deplorable, the necessary repair work cannot be carried out due to a lack of funds.' (corola-56647)

During the data collection and annotation process of intra-sentential anaphoric subjects, we faced a number of problematic cases. Firstly, for both corpora we had to manually extract the complex sentences we were interested in in this study. With the data available to us, we unfortunately had to limit our analysis to a relatively small number of pronominal subjects for the written corpus (68 occurrences). In order to have a balanced set of observations for statistical analyses, we applied the *upSample* function from the *caret* package (Kuhn 2008) that

[6] The person factor was shown to influence the production of referential expressions (see Soares et al. 2020 for Brazilian Portuguese). However, for the goal of this paper, we will only study anaphoric subjects.

allows to add simulated observations without changing the general distribution. Annotating antecedents was constrained at some points by limited access to the previous context (the sentence preceding the target sentence we annotated). This concerned in particular antecedents of pronouns in main clauses and is the main reason why, while we annotated each factor both for main clauses and subordinate clauses, we decided to analyze only the production of null and pronominal subjects occurring in subordinate sentences for the variables where properties of the antecedent were at stake.

2.3 Annotated factors

Comparative studies on pro-drop languages (see Contemori & Di Domenico 2021, Torregrossa et al. 2020, among others) have revealed that crosslinguistic variation may exist in pronoun resolution as well in the choice of referential expressions in production. It is, thus, not necessarily the case that results from previous studies can be taken for granted for Romanian. Following our main research question, we annotated a list of factors which have been shown to affect the choice of antecedents or referential expressions across several languages. Table 1 shows the 15 factors we manually annotated for complex sentences.

Table 1: Factors used for annotation

Annotated element	Factors
Sentence	modality, polarity
Subordinate clause	position (right, left)
Adverbial clause	discourse relation, connectives
Verb	agentivity, mood, tense, voice
Subject	number, gender, animacy, type (null vs. pronominal), place (main vs. subordinate)
Antecedent	syntactic function

While we annotated a variety of factors that may influence the choice of null and pronominal subjects, our general question mainly concerns factors influencing the next mention probability, that means in particular factors affecting the prominence of a referent as well as discourse relations. We generally assume that higher prominence will increase the probability of null subject choices. Following Carminati (2002), we annotated the syntactic function of antecedents in order to test possible preferences for subject antecedents. Animacy has been shown to have a strong influence on pronoun choice in corpus studies on Brazilian Portuguese (Soares et al. 2020, Duarte 2000 a.o.). Agentivity seems to predict preferences for pronoun antecedents aside from topicality and subjecthood in German (Schumacher et al. 2016). Voice also seems to impact pronoun resolution. Colonna et al. (2018) show that the use of passives increases the accessibility of the subject antecedent (see also d'Arcais 1973, Burmester et al. 2018). We predict that the salience-enhancing effect of passives may also increase the probability of null subject choices in production. Moreover, as shown by Rohde & Kehler (2014), pronoun resolution might be sensitive to discourse relations. We therefore also annotated discourse relations. Next mention probability may also be affected by gender, which has been shown to affect choices of subject types in Italian (Cacciari et al. 2011) but also antecedent choices in English (Ferstl et al. 2011). Beyond these factors, verb mood, tense and number have been found to play a role in the frequency of null and pronominal subjects in Granada Spanish (Manjón-Cabeza Cruz et al. 2016).

2.4 Results

All data were analyzed using logistic regressions (*glm* function in the *lme4* package, cf. Bates et al. 2015, *lmerTest* function for *p*-values, cf. Kuznetsova et al. 2017). Our data did not allow us to calculate a single general model with all factors, due to overfitting problems. We therefore analyzed the choice of subject type (dependent variable) based on the annotated factors in the main clause (independent variable) one by one. All factors were mean centered such that *p*-values reflect main effects.[7] Table 2 shows the full set of results in raw numbers. For better comparability with data from previous studies, the variables *Antecedent function*, *Animacy*, *Agentivity*, *Voice*, *Gender*, *Number* and *Discourse relations* reflect subject choices in the subordinate clause. Descriptive (Table 2) and inferential statistics (Table 3) are calculated for subordinate clauses only for these variables. Effects of *Mood*, *Polarity* and *Tense* were calculated across main and subordinate clauses.

Different from previous studies, we did not find significant effects of animacy or number of the main clause subject, or of polarity, mood, and tense.

Figure 1 and Figure 2 show the distribution of null and pronominal subjects in main and subordinate clauses. Null subjects are more frequent in subordinate clauses while pronominal subjects are more frequent in main clauses (Est. = -2.889, std. error = 0.288, z=-10.043, p < 0.001). For both main and subordinate

[7] m = glm(response ~ FactorC, data=data, family = "binomial"), where *response* corresponds to *subject type.*

Table 2: Descriptive statistics (raw numbers)

Factors	Values	Null subjects	Pronominal subjects	Total
Antecedent function	subject	149	31	180
	non-subject	30	30	60
Discourse relations	temporal	57	12	69
	causal	47	48	95
	condition	11	7	18
	concession	18	2	20
	result	14	1	15
	other	35	2	37
Animacy	animate	151	56	207
	non-animate	31	16	47
Agentivity	agentive	93	15	108
	non-agentive	89	57	146
Voice	active	139	63	202
	non-active	43	9	52
Gender (animates)	feminine	43	44	87
	masculine	110	71	181
	other	14	16	30
Number	singular	136	48	184
	plural	46	24	70
Polarity	affirmative	179	158	337
	negative	21	10	31
Mood	indicative	181	153	334
	conditional	4	0	4
	subjunctive	15	15	30
Tense	present	122	109	231
	compound past	57	38	95
	imperfect and pluperfect	10	11	21
	future	10	10	20

Table 3: Inferential statistics

Factors	Estimate	Std. error	z value	Pr (|z|)
Antecedent function	1.570	0.325	4.831	< 0.001
Discourse relations	1.579	0.378	4.177	< 0.001
Animacy	0.058	0.278	0.211	0.832
Voice	-0.773	0.397	-1.947	0.052
Agentivity	1.379	0.326	4.231	< 0.001
Gender (animates)	0.522	0.249	2.097	< 0.05
Number	-0.391	0.303	-1.292	0.197
Polarity	0.617	0.399	1.545	0.122
Mood	0.068	0.362	0.189	0.850
Tense	0.166	0.217	0.767	0.443

clauses, null subjects are more frequent in the written modality (main clauses: Est. = 0.806, std. error = 0.328, z=2.46, $p < 0.05$; subordinate clauses: Est. = -1.333, std. error = 0.333, z=-4.008, $p < 0.001$).

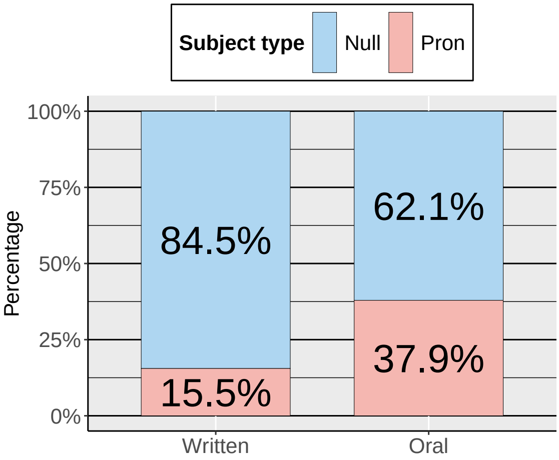

Figure 1: Modality subordinate clauses

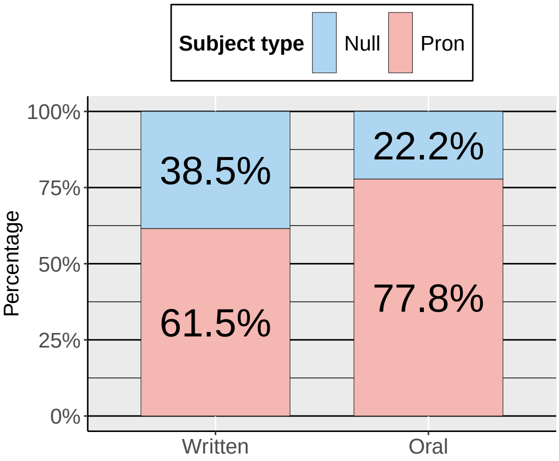

Figure 2: Modality main clauses

For the syntactic function of the antecedent, we only analyzed antecedents for the subordinate clauses. Antecedents for main clauses were not possible to identify in more than 30% of the cases. For the subordinate clauses, we found a significantly higher frequency of null subjects with a subject antecedent (see Table 3 and Figure 3). However, in the case of non-subject antecedents, there is no preference for a null or pronominal subject.

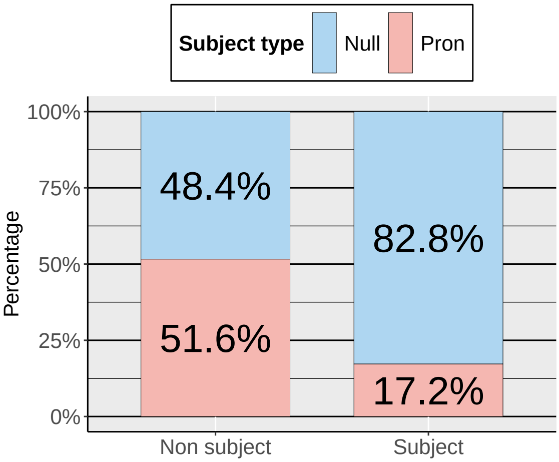

Figure 3: Subject distribution (Antecedent function)

With respect to factors that may increase the salience of an antecedent, we looked at the effects of animacy of the subject and agentivity as well as voice of the verb in the main clause on subject choices in the subordinate clause. Animacy did not show a reliable effect. For voice, we found that non-active voice in the main clause (most often passives) marginally favors null subject pronouns in the upcoming subordinate clause (see Table 3 and Figure 4).

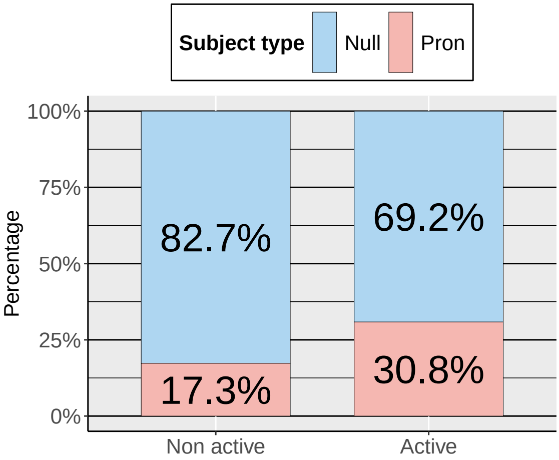

Figure 4: Subject distribution (Voice in main clause)

Another variable we assumed to be related to salience is the agentivity of the verb in the main clause. Agents have a higher next mention probability. In line with this hypothesis, we found that agentive verbs in the main clause lead to significantly more null subjects in the subordinate clause (see Table 3 and Figure 5).

Gender could be a further variable related to antecedent salience. Looking only at animate antecedents, we found that null subjects were reliably more frequent for male antecedents while null and pronominal subjects were equally distributed for female antecedents (see Table 2 and Table 3). This result could be interpreted as evidence that male antecedents are seen as more salient (although more detailed analyses would be necessary to confirm this hypothesis).

Finally, we looked at the distribution of null and overt subjects in the subordinate clauses based on discourse relations (Figure 6). Temporal and causal relations were the most frequent in our data, thus we compared only these two relations through statistical analysis. In adverbial temporal subordinates, we found an increased tendency for producing null subjects compared to pronominal subjects (see Table 3), whereas in causal subordinates the distribution between the two types of subjects is roughly balanced.

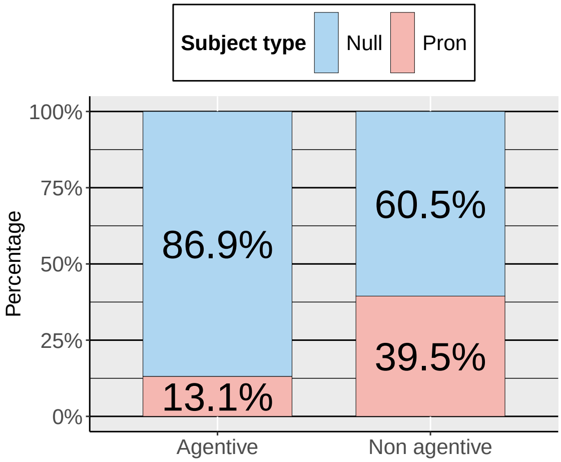

Figure 5: Subject distribution (Agentivity in main clause)

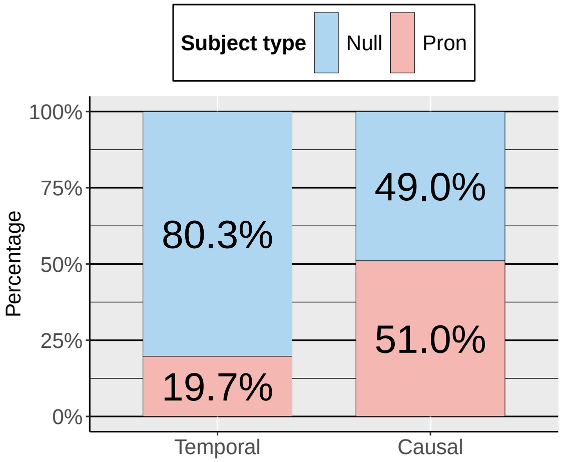

Figure 6: Subject distribution (Discourse relations)

2.5 Discussion: corpus study

Overall, we observed that modality (written vs. spoken) has a clear influence on referential expression production in Romanian. The lower frequency of pronominal subjects in the written corpus might be due to the role of normative grammar reserving the use of pronominal subjects solely for emphatic contexts (cf. Avram 2001). Moreover, we suggest that the higher frequency of pronominal subjects in spoken corpora might be explained by some version of a noisy channel model

(Gibson et al. 2013). According to this model, given the increased noise in oral communication, a rational speaker might opt to be more redundant in order to ensure the message is understood, thus using more pronominal subjects. Regarding antecedent function, null subjects were mostly produced when referring back to a subject antecedent, with an even higher proportion in the written modality. This tendency found for Romanian seems to be in line with previous findings in the literature, postulating that subjects make particular good antecedents (Crawley et al. 1990, Arnold 1998). Further, in a controlled experimental study using items with transitive verbs, Istrate et al. (2022) found a clear preference of null subjects for subject antecedents, although less categorical than in our corpus study (see de la Fuente & Hemforth 2013 for similar tendencies in Spanish). Moreover, in the corpus study, the production of pronominal subjects did not show a clear preference for subject or non-subject antecedents. Grammatical properties of the antecedents, such as *PAH* (Carminati 2002), seem to not fully capture the division of labor in null and pronominal subjects. Moreover, following Givón's (1983) *Topic Continuity Hypothesis*, a preference for null subjects with subject antecedents may be explained by information structure as suggested by Mayol (2010). Subjects as default topics have a higher probability of being mentioned next in the discourse.

Factors influencing the salience of a referent in the discourse, such as agentivity, voice and possibly gender, were shown to play a role in our corpus study. As argued before, we assume that higher salience increases the probability of being mentioned next in the discourse and, thus, the probability of a null subject.

Discourse relations also play a role in the present corpus study. In temporal subordinates, null subjects were more frequent, while for causal relations the results show no clear preference for a referential form. How can these results be related to next mention probabilities or predictability?

The high frequency of null subjects in temporal relations might be explained by the *Topic Continuity Hypothesis* (Givón 1983, Runner & Ibarra 2016). Temporal relations are typically part of a narration where topics rarely shift (see 4a).

For causal relations, we suggest that next mention probabilities may play a role as observed in examples from our corpus data. Implicit causality biases of the verbs may play a role here. In (4b), a null subject in the causal clause retrieves an object antecedent (*doamnei Mihaela* 'to Mrs. Mihaela'). The object in this case is the most predictable referent due to the implicit causality bias of a verb like *a mulțumi* 'thank' that induces an expectation for a reason why Mihaela should be thanked. Null subjects will refer back to the object antecedent when it is foregrounded by the implicit causality bias of the verb. The foregrounded antecedent becomes the most predictable in the discourse. Subject-biased verbs like *fascinate* would predict the subject of the main clause to be mentioned next and to be referred to with a null subject in the causal subordinate clause.

(4) a. Procedând în acest fel, Martin Șluț și-a încălcat
doing in this manner Martin Șluț REFL.DAT.3-AUX.3SG broke.PST
promisiunea făcută anul trecut, atunci când a fost
promise.DEF made.SG.F year.DEF past when AUX.3SG be.PST
ales în fruntea Parlamentului de la Strasburg.
elected in head.DEF parliament.DEF.GEN of Strasbourg
'By doing so, Martin Sluț has broken the promise he made last year when he was elected to lead the Strasbourg Parliament.'

b. Mă numesc Nicolae Maria, doresc să-i
CL.1SG.ACC name.PRS.1SG Nicolae Maria wish.PRS.1SG SBJV-CL.3SG.DAT
mulțumesc doamnei Mihaela pentru că mi-a
thank.PRS.1SG madam.GEN Mihaela because CL.1SG.DAT-AUX.3SG
dezlegat cununia.
save.PST marriage.DEF
'My name is Nicolae Maria, I would like to thank Mihaela because she saved my marriage.' (corola-32168)

The overall results of the corpus study suggest that semantic-pragmatic factors seem to affect the production and interpretation of null and pronominal subjects. Their distribution was shown to be influenced by causal relations which differ considerably compared to temporal relations. However, while the role of implicit causality biases for the choice of null and pronominal subjects in the causal subordinate clause is plausible (Mayol 2018), we cannot confirm it based on the corpus data alone. This is why we ran the controlled experimental study reported in the next section.

3 Experiment: sentence completion task

3.1 Implicit causality verbs

Our corpus study showed no clear preference for null or pronominal subjects for sentences with causal relations. In the following experimental study, we want to shed light on why this may be the case. In previous studies taking into account semantic-pragmatic factors, two main classes of verbs were tested for potential predictability effects: implicit causality verbs (see Caramazza et al. 1977, Costa et al. 2004, Fukumura & Van Gompel 2010, Rohde & Kehler 2014, Holler & Suckow 2016, Mayol 2018, Weatherford & Arnold 2021, Bott & Solstad 2023) and transfer-of-possession verbs (Stevenson et al. 1994, Rohde 2008, Vogels 2019, Lindemann et al. 2020). We will only focus on implicit causality verbs in the following. So-called implicit causality verbs possess an inherent causal meaning introducing a semantic bias towards continuations referring back to the entity related to the underlying causer of the event, which can appear in either subject position (subject-biased implicit causality verbs) or object position (object-biased implicit causality verbs). In (5a), the verb bias increases the prediction of the upcoming cause to be attributed to the subject *Mary* while in (5b) the upcoming cause is predicted to be attributed to the object *Peter*.

(5) a. *Mary* fascinated Peter because ...→ *Mary* more likely continuation
b. Mary criticized *Peter* because ...→ *Peter* more likely continuation

Ferstl et al. (2011) moreover suggest that, beyond the next mention bias invoked by the verb, the gender of the antecedents may play a role in that male antecedents have a slightly higher probability of being seen as the causer of an event. The general gender effect we found in our corpus study makes a similar prediction.

3.2 Methods

3.2.1 Hypotheses

With a sentence completion task, we tried to answer the following hypotheses that are, to some extent, interconnected. According to the *PAH* (Carminati 2002, and see also *topic continuity* in Givón 1983 or similar approaches), null subjects have a strong tendency to go with a subject antecedent, which is compatible with the corpus study presented in the previous section. This hypothesis predicts that participants produce more null subjects when referring back to a subject antecedent. Pronominal subjects should be used more when participants refer back to non-subjects. If, however, null subjects prefer more predictable antecedents that are likely to be mentioned next, verb biases may change the picture: Null subjects should be more frequent when the continuation is in line with the verb bias.

The gender of the antecedents might also be affected by next mention biases, as suggested by Ferstl et al. (2011). Our results from the previous corpus study revealed that masculine antecedents are more prominent, thus we also expect to observe a higher preference for null subjects with masculine antecedents compared to feminine ones.

To sum up our hypotheses:

- A null subject will be produced more often when retrieving a subject antecedent.
- Based on the implicit causality biases of the verbs, continuations should refer to the antecedent foregrounded by the verb (the more predictable antecedent). The choice of the referential expression (null vs. pronominal subject) will then be influenced by the verb bias.
- Following results from Ferstl et al. (2011), we may also find a gender effect with a preference to choose male antecedents as the causer of an event, leading to a preference for null subjects to refer to male antecedents.

3.2.2 Participants

Thirty-one native Romanian speakers (age range 19 to 32 years, mean age: 27 years) participated in our experiment. All of the participants spent their childhood in Romania. They were students recruited at the University of Bucharest. Given that the participants are enrolled in an institution of higher education, their level of instruction is fairly homogeneous (a minimum of 12 years of instruction). Thus, the participants had no difficulty in reading, understanding, or continuing the sentences. Participation was voluntary and participants were not paid for their contribution. The experiment was run on a version of Ibex farm installed on a local server at Université Paris Cité. Participants' data were immediately anonymized. At no moment was identifying information stored.

3.2.3 Materials

The experiment focuses on testing the production of referential expressions (lexical vs. pronominal vs. null subject) in Romanian as well as the preference for an antecedent using a free passage completion task with a paradigm similar to Kehler & Rohde (2019). In order to increase the predictability of an antecedent, we chose implicit causality verbs that increase the next mention probability of the subject as in (6a)-(6b) with subject-biased verbs, or the object as in (6c)-(6d) with object-biased verbs.

(6) a. Maria îl dezamăgește pe Victor pentru că ...
Maria CL.3SG.M.ACC disappoint.PRS.3SG DOM Victor because
'Maria disappoints Victor because...'

b. Victor o dezamăgește pe Maria pentru că …
Victor cl.3sg.f.acc disappoint.prs.3sg dom Maria because
'Victor disappoints Maria because …'

c. Alexandra îl adoră pe Albert pentru că …
Alexandra cl.3sg.m.acc adore.prs.3sg dom Albert because
'Alexandra adores Albert because …'

d. Albert o adoră pe Alexandra pentru că …
Albert cl.3sg.f.acc adore.prs.3sg dom Alexandra because
'Albert adores Alexandra because …'

One of the antecedents was always a feminine first name, the other a masculine first name. We created two conditions for each sentence switching the gender of the subject and the object to test for possible gender effects as they were found in Ferstl et al. (2011). Participants were asked to continue sentences following the pattern in (7) using a plausible continuation of their choice (freely choosing a referential expression for the subject of the causal clause). According to the literature, participants continue with either a pronominal, a lexical or a null subject in more than 85% of the cases (see Kehler & Rohde 2019).

(7) Female/Male first name + implicit causality subject/object-bias verb + Male/Female first name + *because*

In order to create our experimental items, we selected a total of 48 implicit causality verbs (24 subject-biased verbs and 24 object-biased verbs) chosen from the database created by Ferstl et al. (2011). Given the fact that there is no similar database in Romanian, we based the choice of our verbs on the English verbs with the highest implicit causality biases (above 70%) which were then translated and adapted to Romanian. The verbs as well as the experimental items were reviewed by an independent native Romanian speaker (other than the creators and annotators of the experiment). We selected approximately 50% of verbs with a positive connotation (e.g. *impress, congratulate*) and 50% with a negative connotation (e.g. *disappoint, envy*). The names used for our items were very common, well-known, typical Romanian names to limit other potential biases as much as possible.

3.3 Procedure

Completions in the kind of task we use here can be free or constrained. In a constrained completion task, participants are invited to write completions referring back to an entity that is somehow marked. For the unconstrained or free sentence completion task that we applied, participants were asked to complete the items without any constraints for the antecedent, providing likely continuations to the given sentences. Relying on the strength of the implicit causality biases, we opted for a free completion task (for a discussion of advantages and disadvantages of both paradigms, see Demberg et al. 2023). The task was conducted online on the Ibex Farm platform at Université Paris Cité (created by Alex Drummond and maintained by Achille Falaise). The experiment began with instructions for the task as well as a series of demographic questions (age, gender, first language, i.e. language spoken since early childhood). Participants gave their informed consent to the use of their anonymized data for research purposes. A total of 1071 completions were annotated excluding ambiguous or inappropriate answers. The continuations were independently annotated separately by two native Romanian speakers (both coauthors of the paper) with respect to the intended antecedent of the continuation as well as with respect to the referential expression used for the subject of the causal subordinate sentence introduced by *because*.

3.4 Results

In the annotation process, the antecedent of a null subject cannot be determined by syntactic markers given the nature of a null pronoun. Hence, when null subjects were produced, antecedent choice was determined by the meaning of the causal subordinate clause. The two annotators agreed on all decisions.[8] All data were analyzed using logistic regressions (*glmer* function in the *lme4* package, cf. Bates et al. 2015, *p*-value being estimated using *lmerTest*, cf. Kuznetsova et al. 2017). We first analyzed the effect of implicit causality and gender on antecedent choice. Gender of the subject of the root clause as well as verb bias were added as mean centered fixed factors and participants and items as random factors. Random slopes could not be added due to convergence failure. This is true for all models presented here. As shown in Figure 7, participants' continuations were highly consistent with the verb bias. They mostly chose a continuation consistent with a subject antecedent after subject-biased verbs and with the object antecedent after object-biased verbs (Est. = -6.90, std. error = .6176, z=-11.179, p <.001). There was also a small numeric effect of gender with slightly more subject choices (less object choices) when the subject antecedent was male (Est. = -.5938, std. error = .3586, z=-1.656, p = .0977).

[8] E.g., in a sentence like *Peter thanked David because he proofread the thesis*, it is highly plausible to assume that the null subject refers back to the object of thanking.

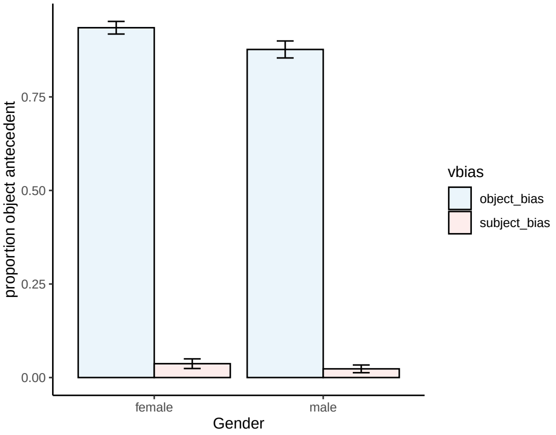

Figure 7: Next mention and verb bias

Participants chose null or pronominal subjects in more than 95% of the cases. We therefore excluded other referential expressions from our analyses. Participants overwhelmingly produced null subjects in the *because*-subordinate clause independently of verb bias (Est. = 4.8020, std. error = .6262, z=7.669, p < .001). As shown in Figure 8, null subjects were moreover chosen more frequently for sentences with subject-biased verbs where verb bias and the general preference of null subjects for subject antecedents align (Est. = 1.6934, std. error = .5866, z=2.887, p <.01). We finally looked at the frequencies of null and pronominal subjects depending on the antecedent choice made by the participants. Logistic regressions included antecedent choice and verb bias as fixed factors (both mean centered) and participants and items as random factors. Figure 9 shows that participants chose null subjects more often in cases where the verb bias and the antecedent choice aligned, i.e. when they produced a continuation consistent with an object antecedent in sentences with object-biased verbs or a continuation consistent with a subject antecedent in sentences with subject-biased verbs (Est. = 4.3433, std. error = 1.2946, z=3.355, p < .001).

3.5 Discussion

All in all, the results of our free passage completion task in Romanian confirm the hypothesis that null subjects are the most preferred referential form to retrieve subject antecedents. However, null subjects can also easily be produced for non-subject antecedents when they are highly predictable in the context. This can

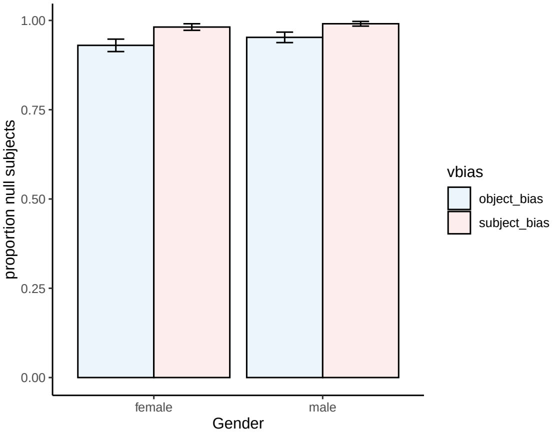

Figure 8: Subject antecedents and verb bias

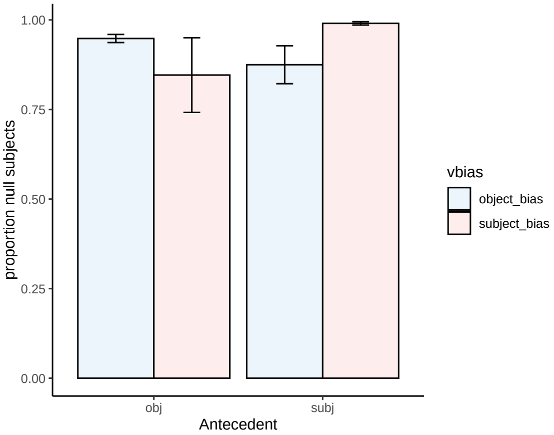

Figure 9: Subject choice and verb bias

be interpreted as is clear evidence against at least a simple version of the *PAH* (Carminati 2002) that stipulates a general preference of null subject pronouns for subject antecedents. While it might be argued that the experimental situation could lead to non-natural productions from the participants, the continuations produced by the participants with object antecedents in the case of object-biased verbs (see 8a and 8b for examples from the experiment) were judged as highly natural by both native speakers of Romanian who annotated them.

(8) a. Laura îl felicită pe Ionuț pentru că
Laura CL.3SG.M.ACC congratulate.PRS.3SG DOM Ionuț because
a luat permisul.
AUX.3SG take.PST license.DEF.M.SG
'Laura congratulates Ionuț because he got his driving license.'

b. Ionuț o felicită pe Laura pentru că
Ionuț CL.3SG.F.ACC congratulate.PRS.3SG DOM Laura because
a câștigat un permiu.
AUX.3SG win.PST a prize
'Ionuț congratulates Laura because she won a prize.'

In an experimental study on pronoun choice and thematic roles, Lindemann et al. (2020) suggest a very similar pattern for Romanian. Their results also show a general preference of null subjects for subject antecedents with no clear preference for pronominal subjects. Interestingly, in their study, thematic roles (goal vs. source) affected the production of referring expressions alongside the grammatical role of the antecedents. Null subjects were more often used to retrieve *goal* referents, i.e. more prominent or predictable referents.

In our experiment, we observe that participants produced pronominal subjects more often when the antecedent was less salient or predictable. However, other cases of pronominal subjects (as in 9a and 9b) included continuations that contained a contrast between antecedents (see Dobrovie-Sorin & Giurgea 2013 and Mayol 2010 for similar suggestions). The role of contrast has more recently also been confirmed in experimental studies by Istrate et al. (2024).

(9) a. Maria îl dezamăgește pe Victor pentru că și el
Maria CL.3SG.M.ACC disappoint.PRS.3SG DOM Victor because also he
a făcut același lucru.
AUX.3SG do.PST same thing
'Maria disappoints Victor because he also did the same thing.'

b. Victor o invidiază pe Maria pentru că ea are note
Victor CL.3SG.F.ACC envy.PRS.3SG DOM Maria because she has grades
mai bune decât el.
better than him
'Victor envies Mary because she has better grades than him.'

With respect to the predictability effect induced by implicit causality verbs, tendencies in Romanian are fairly similar to those suggested by Bott & Solstad

(2023) for German in which the referent predictability was shown to have a strong impact on pronoun production. Moreover, from a crosslinguistic perspective, Romanian seems to align with preferences found for other pro-drop Romance languages (see Contemori & Di Domenico 2021 for Italian and Spanish, Mayol 2018 for Catalan), i.e. null subject pronouns will be favoured for more predictable referents. For European Portuguese, Costa et al. (2004) showed similar preferences for null subjects, but a different pattern in the case of pronominal subjects, which were used more often for object antecedents foregrounded by an object-biased verb. Regarding gender biases, we found a small numeric gender effect (following Ferstl et al. 2011) in our experimental data.

Unlike Bott & Solstad (2023) and Rosa & Arnold (2017), who suggest that predictability effects are stronger with same-gender antecedents, we found that next-mention bias can also play a strong role when using different-gender antecedents in an implicit causality experiment.

4 General discussion

In our corpus study, we found that, while null subjects have a strong preference for subject antecedents as predicted by the *PAH* (Carminati 2002), other more semantic-pragmatic factors also play a role. In particular, prominence enhancing factors such as *voice* or *agentivity* (and potentially *gender*) but also *modality* affect the choice of the referential expression. Moreover, discourse relations seem to play a role in that null and pronominal subjects are equally distributed in causal relations but not in temporal relations.

In our experimental study, we tried to better understand the specific pattern for causal relations. Despite a slight general preference for subject antecedents, null subjects were shown to be strongly preferred as referential form for non-subject antecedents as well when they were predictable enough in the context. More concretely, object antecedents were retrieved mostly by a null subject in the context of causal coherence relations, with an implicit causality verb biased towards the object (e.g., *congratulate*). While null subjects were preferred when the continuations aligned with the verb bias, pronominal subjects were generally used by participants in continuations which go against the verb bias (for example, in contrastive contexts; for similar suggestions see Mayol 2010).

Moreover, putting our corpus study and experimental results together, we contribute to a broader research question which is under significant debate, more specifically, whether predictability influences the choice of referring expressions (cf. Arnold 2001, Fukumura & Van Gompel 2010, Rohde & Kehler 2014, Holler &

Suckow 2016, Modi et al. 2017, Rosa & Arnold 2017 a.o.). Both corpus and experimental evidence in Romanian suggest that higher predictability (Tily & Piantadosi 2009) triggers a clear preference for null subject pronouns. This effect of referent predictability is also in line with previous hypotheses in the literature on the role of salience or accessibility (Givón 1983, Ariel 1994, Grosz et al. 1995, Chafe 1996) or prominence (von Heusinger & Schumacher 2019). As suggested by Demberg et al. (2023), the next mention bias may be triggered by a complex interaction of two semantic-pragmatic factors, i.e. the verb bias and coherence relations.

In general, our data on Romanian replicate results from previous studies (Costa et al. 2004, Mayol 2018, Contemori & Di Domenico 2021) on languages from the Romance family, while they are inconsistent with previous data from Mandarin Chinese (Hwang et al. 2022). More crosslinguistic studies are needed to establish in how far general pro-drop patterns in a language may play a role here.

Data and materials are accessible here: https://osf.io/fmjnq/.

Acknowledgements

We want to thank Anne Abeillé and Gabriela Bîlbîie for their ideas concerning the annotated factors in the corpus study. We also thank Verginica Barbu-Mititelu for her assistance in collecting data from Romanian corpora. And finally we would like to thank Labex EFL and Smarts-Up (Idex Université Paris Cité) for co-funding our work.

References

Ariel, Mira. 1994. Interpreting anaphoric expressions: A cognitive versus a pragmatic approach. *Journal of linguistics* 30(1). 3–42. DOI: https://doi.org/10.1017/S0022226700016170.

Arnold, Jennifer E. 1998. *Reference form and discourse patterns*. Stanford University.

Arnold, Jennifer E. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse processes* 31(2). 137–162. DOI: 10.1207/S15326950DP3102_02.

Arnold, Jennifer E. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4(4). 187–203. DOI: https://doi.org/10.1111/j.1749-818X.2010.00193.x.

Avram, Mioara. 2001. *Gramatica pentru toţi*. 2nd edn. Bucureşti: Humanitas.

Barbu Mititelu, Verginica, Dan Tufiş & Elena Irimia. 2018. The reference corpus of the contemporary Romanian language (CoRoLa). In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1178–1185. European Language Resources Association (ELRA).

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using 'lme4'. *Journal of Statistical Software* 67(1). 1–48. DOI: 10.18637/jss.v067.i01.

Bott, Oliver & Torgrim Solstad. 2023. The production of referring expressions is influenced by the likelihood of next mention. *Quarterly Journal of Experimental Psychology* 76(10). 2256–2284. DOI: 10.1177/17470218231157268.

Burmester, Juliane, Antje Sauermann, Katharina Spalek & Isabell Wartenburger. 2018. Sensitivity to salience: Linguistic vs. visual cues affect sentence processing and pronoun resolution. *Language, Cognition and Neuroscience* 33(6). 784–801. DOI: 10.1080/23273798.2018.1428758.

Cacciari, Cristina, Paola Corradini, Roberto Padovani & Manuel Carreiras. 2011. Pronoun resolution in Italian: The role of grammatical gender and context. *Journal of Cognitive Psychology* 23(4). 416–434. DOI: 10.1080/20445911.2011.526599.

Caramazza, Alfonso, Ellen Grober, Catherine Garvey & Jack Yates. 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior* 16(5). 601–609. DOI: https://doi.org/10.1016/S0022-5371(77)80022-4.

Carminati, Maria Nella. 2002. *The processing of Italian subject pronouns*. University of Massachusetts Amherst.

Chafe, Wallace. 1996. Inferring identifiability and accessibility. In Thorstein Fretheim & Jeanette K. Gundel (eds.), *Reference and referent accessibility*, 37–46. Amsterdam: John Benjamins Publishing Company.

Chamorro, Gloria. 2018. Offline interpretation of subject pronouns by native speakers of Spanish. *Glossa: a journal of general linguistics* 3(1). DOI: https://doi.org/10.5334/gjgl.256.

Colonna, Saveria, Sarah Schimke, Israel de la Fuente, Sascha Kuck & Barbara Hemforth. 2018. Effects of exposure and information structure in native and non-native pronoun resolution in French. *Linguistics Vanguard* 4(S1). 20160093. DOI: 10.1515/lingvan-2016-0093.

Colonna, Saveria, Sarah Schimke & Barbara Hemforth. 2012. Information structure effects on anaphora resolution in German and French: A crosslinguistic study of pronoun resolution. *Linguistics* 50(5). 991–1013. DOI: https://doi.org/10.1515/ling-2012-0031.

Contemori, Carla & Elisa Di Domenico. 2021. Microvariation in the division of labor between null-and overt-subject pronouns: The case of Italian and Spanish. *Applied Psycholinguistics* 42(4). 997–1028. DOI: https://doi.org/10.1017/S0142716421000199.

Costa, Armanda, Isabel Hub Faria & Michèle Kail. 2004. Semantic and syntactic cues' interaction on pronoun resolution in European Portuguese. In António Branco, Tony McEnery & Ruslan Mitkov (eds.), *DAARC 2004, Proceedings of 5th Discourse Anaphora Resolution Colloquium*, 45–50. Lisboa: Ed. Colibri.

Crawley, Rosalind A., Rosemary J. Stevenson & David Kleinman. 1990. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research* 19. 245–264. DOI: 10.1007/BF01077259.

d'Arcais, Giovanni B. Flores. 1973. *Some perceptual determinants of sentence construction.* University of Leiden.

de la Fuente, Israel & Barbara Hemforth. 2013. Effects of clefting and left-dislocation on subject and object pronoun resolution in Spanish. In Jennifer Cabrelli Amaro, Gillian Lord, Ana de Prada Pérez & Jessi Elana Aaron (eds.), *Selected Proceedings of the 16th Hispanic Linguistics Symposium*, 27–45. Somerville, MA: Cascadilla Proceedings Project.

Demberg, Vera, Ekaterina Kravtchenko & Jia E. Loy. 2023. A systematic evaluation of factors affecting referring expression choice in passage completion tasks. *Journal of Memory and Language* 130. 104413. DOI: 10.1016/j.jml.2023.104413.

Dobrovie-Sorin, Carmen & Ion Giurgea. 2013. *A reference grammar of Romanian: Volume 1: The noun phrase*, vol. 207. John Benjamins Publishing. DOI: https://doi.org/10.1075/la.207.

Duarte, Maria Eugênia Lamoglia. 1995. *A perda do princípio 'evite pronome' no português brasileiro.* Brazil: Universidade Estadual de Campinas. (Doctoral dissertation).

Duarte, Maria Eugênia Lamoglia. 2000. The loss of the 'avoid pronoun' principle in Brazilian Portuguese. In Mary Aizawa Kato & Esmeralda Vailati Negrão (eds.), *Brazilian Portuguese and the null subject parameter*, 17–36. Frankfurt a. M./Madrid: Vervuert Verlagsgesellschaft.

Ferretti, Todd R., Hannah Rohde, Andrew Kehler & Melanie Crutchley. 2009. Verb aspect, event structure, and coreferential processing. *Journal of Memory and Language* 61(2). 191–205. DOI: https://doi.org/10.1016/j.jml.2009.04.001.

Ferstl, Evelyn C., Alan Garnham & Christina Manouilidou. 2011. Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods* 43. 124–135. DOI: https://doi.org/10.3758/s13428-010-0023-2.

Fukumura, Kumiko, Coralie Hervé, Sandra Villata, Shi Zhang & Francesca Foppolo. 2022. Representations underlying pronoun choice in Italian and English. *Quarterly Journal of Experimental Psychology* 75(8). 1428–1447. DOI: 10.1177/17470218211051989.

Fukumura, Kumiko & Roger P.G. Van Gompel. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language* 62(1). 52–66. DOI: https://doi.org/10.1016/j.jml.2009.09.001.

Fukumura, Kumiko & Roger P.G. Van Gompel. 2011. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes* 26(10). 1472–1504. DOI: https://doi.org/10.1080/01690965.2010.506444.

Gernsbacher, Morton Ann & David J. Hargreaves. 1988. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language* 27(6). 699–717. DOI: https://doi.org/10.1016/0749-596X(88)90016-2.

Gibson, Edward, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim & Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological science* 24(7). 1079–1088. DOI: 10.1177/0956797612463705.

Givón, Talmy. 1983. Topic continuity in discourse: The functional domain of switch reference. In John Haiman & Pamela Munro (eds.), *Switch reference and universal grammar*, 51–82. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/tsl.2.06giv.

Gordon, Peter C., Barbara J. Grosz & Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive science* 17(3). 311–347. DOI: https://doi.org/10.1207/s15516709cog1703_1.

Grice, Herbert P. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and Semantics: Speech Acts*, 41–58. New York: Academic Press.

Grosz, Barbara, Aravind Joshi & Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21(2). 203–225.

Himmelmann, Nikolaus P. & Beatrice Primus. 2015. Prominence beyond prosody: A first approximation. In Amedo De Dominicis (ed.), *Prominence in linguistics. Proceedings of the pS-prominences International Conference*, 38–58. University of Tuscia.

Holler, Anke & Katja Suckow. 2016. *Empirical perspectives on anaphora resolution*, vol. 563. Walter de Gruyter GmbH & Co KG. DOI: https://doi.org/10.1515/9783110464108.

Holmberg, Anders. 2010. Null subject parameters. In Theresa Biberauer, Anders Holmberg, Roberts Ian & Michelle Sheehan (eds.), *Parametric variation: Null subjects in minimalist theory*, 88–124. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511770784.003.

Hwang, Heeju, Suet Ying Lam, Wenjing Ni & He Ren. 2022. The role of grammatical role and thematic role predictability in reference form production in Mandarin Chinese. *Frontiers in Psychology* 13. 930572. DOI: 10.3389/fpsyg.2022.930572.

Istrate, Fabian, Anne Abeillé & Barbara Hemforth. 2022. The position of antecedent hypothesis in Romanian subject alternation: Two experiments. In *Going Romance 2022*. Barcelona.

Istrate, Fabian, Anne Abeillé & Barbara Hemforth. 2024. Subject alternation and antecedent preference in Romanian. *Discours. Revue de linguistique, psycholinguistique et informatique* 34.

Kehler, Andrew & Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics* 39(1-2). 1–37. DOI: https://doi.org/10.1515/tl-2013-0001.

Kehler, Andrew & Hannah Rohde. 2019. Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics* 154. 63–78. DOI: https://doi.org/10.1016/j.pragma.2018.04.006.

Kuhn, Max. 2008. Building predictive models in R using the 'caret' package. *Journal of Statistical Software* 28(5). 1–26. DOI: 10.18637/jss.v028.i05.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune Haubo Bojesen Christensen. 2017. 'lmerTest' package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). DOI: 10.18637/jss.v082.i13.

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John Platt & Thomas Hofmann (eds.), *Advances in neural information processing systems 19: Proceedings of the 2006 conference.* MIT Press. DOI: https://doi.org/10.7551/mitpress/7503.003.0111.

Lindemann, Sofiana-Iulia, Stanca Mada, Laura Sasu & Madalina Matei. 2020. Thematic role and grammatical function affect pronoun production. *ExLing 2020.* 113. DOI: 10.36505/ExLing-2020/11/0028/000443.

Manjón-Cabeza Cruz, Antonio, Francisca Pose Furest & Francisco José Sánchez García. 2016. Factores determinantes en la expresión del sujeto pronominal en el corpus PRESEEA de Granada. *Boletín de filología* 51(2). 181–207. DOI: 10.4067/S0718-93032016000200007.

Mayol, Laia. 2010. Refining salience and the position of antecedent hypothesis: A study of Catalan pronouns. In *University of Pennsylvania Working Papers in Linguistics*. University of Pennsylvania.

Mayol, Laia. 2018. Asymmetries between interpretation and production in Catalan pronouns. *Dialogue & Discourse* 9(2). 1–34. DOI: https://doi.org/10.5087/dad.2018.201.

Modi, Ashutosh, Ivan Titov, Vera Demberg, Asad Sayeed & Manfred Pinkal. 2017. Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics* 5. 31–44. DOI: 10.1162/tacl_a_00044.

Oakhill, Jane, Alan Garnham & Wietske Vonk. 1989. The on-line construction of discourse models. *Language and cognitive processes* 4(3-4). SI263–SI286. DOI: 10.1080/01690968908406370.

Patterson, Clare, Petra B. Schumacher, Bruno Nicenboim, Johannes Hagen & Andrew Kehler. 2022. A Bayesian approach to German personal and demonstrative pronouns. *Frontiers in psychology* 12. 672927. DOI: https://doi.org/10.3389/fpsyg.2021.672927.

Rohde, Hannah. 2008. *Coherence-driven effects in sentence and discourse processing*. University of California, San Diego. (Doctoral dissertation).

Rohde, Hannah & Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience* 29(8). 912–927. DOI: https://doi.org/10.1080/01690965.2013.854918.

Rosa, Elise C. 2015. *Semantic role predictability affects referential form*. The University of North Carolina at Chapel Hill. (Doctoral dissertation). DOI: https://doi.org/10.17615/5w75-3x78.

Rosa, Elise C. & Jennifer E. Arnold. 2017. Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language* 94. 43–60. DOI: 10.1016/j.jml.2016.07.007.

Runner, Jeffrey T. & Alyssa Ibarra. 2016. Information structure effects on null and overt subject comprehension in Spanish. In Anke Holler & Katja Suckow (eds.), *Empirical perspectives on anaphora resolution*, 87–112. Berlin: De Gruyter. DOI: 10.1515/9783110464108-006.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartin, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press.

Schäfer, Lisa. 2021. Topic drop in German: Empirical support for an information-theoretic account to a long-known omission phenomenon. *Zeitschrift für Sprachwissenschaft* 40(2). 161–197. DOI: 10.1515/zfs-2021-2024.

Schulz, Miriam, Heather Burnett & Barbara Hemforth. 2021. Corpus, experimental and modeling investigations of cross-linguistic differences in pronoun resolution preferences. *Glossa: A journal of general linguistics* 6(1). DOI: https://doi.org/10.5334/gjgl.1142.

Schumacher, Petra B., Manuel Dangl & Elyesa Uzun. 2016. Thematic role as prominence cue during pronoun resolution in German. In Anke Holler & Katja Suckow (eds.), *Empirical perspectives on anaphora resolution*, vol. 121, 213–240. Berlin: De Gruyter. DOI: 10.1515/9783110464108-011.

Soares, Eduardo Correa, Philip Miller & Barbara Hemforth. 2020. The effect of semantic and discourse features on the use of null and overt subjects: A quantitative study of third person subjects in Brazilian Portuguese. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 36(1). 2020360107. DOI: 10.1590/1678-460x2020360107.

Stevenson, Rosemary J., Rosalind A. Crawley & David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and cognitive processes* 9(4). 519–548. DOI: https://doi.org/10.1080/01690969408402130.

Tily, Harry & Steven Piantadosi. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. In Kees van Deemter, Albert Gatt, R. van Gompel & Emiel Krahmer (eds.), *Proceedings of the Workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference.* Tilburg University.

Torregrossa, Jacopo, Maria Andreou & Christiane M. Bongartz. 2020. Variation in the use and interpretation of null subjects: A view from Greek and Italian. *Glossa: A journal of general linguistics* 5(1). DOI: doi.org/10.5334/gjgl.1011.

Vogels, Jorrig. 2019. Both thematic role and next-mention biases affect pronoun use in Dutch. In Ashok Goel, Colleen Seifert & Christian Freksa (eds.), *Proceedings of the 41st Annual Conference of Cognitive Science Society*, 3029–3035. Cognitive Science Society. DOI: 10.34894/cytanw.

von Heusinger, Klaus & Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics* 154. 117–127. DOI: https://doi.org/10.1016/j.pragma.2019.07.025.

Weatherford, Kathryn C. & Jennifer E. Arnold. 2021. Semantic predictability of implicit causality can affect referential form choice. *Cognition* 214. 104759. DOI: https://doi.org/10.1016/j.cognition.2021.104759.