DOI: 10.31178/RCSDLLF.12.5

L'alternance entre le sujet nul et pronominal en roumain dans une perspective romane

FABIAN ISTRATE[1]
Sous la direction d'ANNE ABEILLÉ, GABRIELA BÎLBÎIE et BARBARA HEMFORTH

Abstract

This article focuses on the differences in the comprehension and the production of null and pronominal subjects in Romanian. It shows the importance of an empirical approach (experiments and corpus investigation) for the study of syntactic alternations, such us the null pronominal subject alternation in prodrop Romance languages, which cannot be described by the classical categorical syntactic constraints (cf. *grammatical* vs. *agrammatical* dychotomy), but rather by soft constraints, that reflect the preferences of the speakers for one structure or the other. This research allows us to observe (micro)variations within prodrop Romance languages. It also shows that an adequate analysis of such an alternation should be based on a multifactorial analysis, that takes into account factors from various linguistic levels (in particular syntax and discourse).

Keywords: null subject, pronominal subject, alternation, antecedent, Romance, empirical approach

1. Introduction

1.1. L'alternance sujet nul / sujet pronominal et la résolution des pronoms

Dans les langues dites à sujet nul (appelées aussi langues *prodrop*) comme le roumain, qui n'exigent pas la réalisation phonologique du sujet (Dobrovie-Sorin et Giurgea 2013), on observe l'alternance syntaxique entre deux structures possibles, à savoir l'alternance entre une phrase à sujet nul comme en (1a), et une phrase à sujet pronominal comme en (1b).

(1) a. ***Ø*** *Este fratele meu mai mare.*
'Il est mon frère aîné.'

[1] Master de Sciences du Langage, parcours Linguistique Théorique et Expérimentale, Université Paris Cité, France.

b. ***El*** *este fratele meu mai mare.*
'Il est mon frère aîné.'

Dans les travaux traditionnels sur le roumain, les ouvrages de référence (par exemple *GALR* 2005) recommandent de manière générale l'emploi du sujet nul, l'usage des sujets pronominaux étant réservé seulement aux contextes d'emphase ou de contraste. De ce fait, la co-occurrence d'un pronom sujet réalisé et d'un verbe (dont les désinences donnent des indices sur la personne et le nombre du référent) est critiquée par la norme et jugée redondante (*GALR* 2005, Pană-Dindelegan 2003, 2013).

Les recherches récentes dédiées aux alternances syntaxiques (Bresnan *et al.* 2010) suggèrent l'existence des contraintes non-catégoriques (*soft constraints*), appelées aussi contraintes préférentielles (Keller 2000, Thuilier 2012), qui décrivent les préférences manifestées par les locuteurs d'une langue lorsqu'ils ont le choix entre deux structures grammaticales. Pour étudier les alternances syntaxiques, on doit donc aller au-delà des contraintes catégoriques, binaires (décrites par la distinction classique *grammatical* vs *agrammatical*, appelées aussi *hard constraints* en anglais) et observer les contraintes préférentielles, non-catégoriques, qui favorisent, ou non, l'usage d'une forme ou d'une construction (Goldberg 1995). Les préférences des locuteurs peuvent être étudiées et mesurées grâce à l'appel aux données quantitatives, à savoir les études expérimentales et les études de corpus.

Le champ de recherche dans lequel nous nous situons est donc celui de la syntaxe quantitative et expérimentale, issue de la nécessité de prendre en compte la nature graduée des données langagières, afin d'étudier les aspects cognitifs et fonctionnels qui participent à la production et à la perception du langage (Thuilier 2012, Bîlbîie *et al.* 2021).

L'alternance sujet nul / sujet pronominal a fait l'objet d'une attention particulière en linguistique expérimentale (Carminati 2002, Chamorro 2018, Contemori et Di Domenico 2021, Torregrossa *et al.* 2020), car elle constitue l'un des phénomènes syntaxiques majeurs qui permettent l'étude de la résolution anaphorique, c'est-à-dire la manière dont les expressions anaphoriques (en particulier, les pronoms) récupèrent leurs antécédents dans le discours. On peut donc dire que l'étude de cette alternance syntaxique va au-delà du champ de recherche syntaxique et touche à d'autres domaines, comme la sémantique et la pragmatique, allant jusqu'à la psycholinguistique (Holler et Suckow 2016, Arnold *et al.* 2000, Kehler et Rohde 2019).

1.2. Hypothèses avancées dans la littérature

L'alternance sujet nul / sujet pronominal a été étudiée dans les langues romanes prodrop tant en corpus qu'à travers l'expérimentation linguistique, et

plusieurs hypothèses qui pourraient rendre compte de cette alternance ont été avancées dans la littérature.

Un patron syntaxique assez étudié dans les études expérimentales sur les langues romanes se retrouve dans les exemples de (2) à (5) : (2) pour l'italien, (3) pour l'espagnol, (4) pour le portugais et (5) pour le catalan. Ainsi, lorsqu'une phrase principale contenant un sujet et un objet (tous deux désignant des entités de même genre, susceptibles de créer une ambiguïté d'interprétation) enchaîne sur une deuxième phrase (en général, une subordonnée temporelle) qui commence par un sujet anaphorique (soit sujet nul – non réalisé lexicalement, soit sujet pronominal), l'interprétation du sujet anaphorique semble suivre certaines tendances, décrites par l'hypothèse de la position de l'antécédent (*Position of Antecedent Hypothesis*, v. Carminati 2002, 2005). Cette hypothèse prédit que : (i) les sujets nuls préfèrent un antécédent sujet et (ii) les sujets pronominaux préfèrent un antécédent non-sujet. Ainsi, dans un exemple comme (2) en italien, où et le sujet et l'objet de la phrase principale ont des prénoms féminins, les locuteurs auraient tendance à utiliser le sujet nul dans la subordonnée s'il réfère à l'antécédent sujet (*Ø = Flavia*) et le sujet pronominal s'il réfère à l'antécédent objet (*lei = Paola*).

(2) [**italien**, v. Carminati 2002 : 378]
*Flavia ha telefonato a Paola quando **Ø/lei** era in ufficio.*
'Flavia a téléphoné à Paola quand elle était dans le bureau.'

(3) [**espagnol**, v. de la Fuente et Hemforth 2013 : 3]
*Juan vio a Pedro cuando **Ø/él** caminaba por la playa.*
'Juan a vu Pedro quand il se promenait sur la plage.'

(4) [**portugais**, v. Fernandes *et al.* 2018 : 5]
*O atleta consultou o ortopedista quando **Ø/ele** regressou da viagem a Italia.*
'L'athlète a consulté l'orthopédiste quand il est revenu de son voyage en Italie.'

(5) [**catalan**, v. Mayol 2010 : 3]
*La Marta escrivia sovinta la Raquel. **Ø/ella** vivia als Estats Units*.
'Marta écrivait fréquemment à Raquel. Elle vivait aux États-Unis.'

Une deuxième hypothèse que l'on peut trouver dans la littérature pour expliquer la résolution des expressions anaphoriques dans les cas ambigus, mettant en jeu une alternance potentielle entre un sujet nul et un sujet pronominal, remonte à l'étude de Givón 1983, qui postule l'hypothèse de la continuité du topique, selon laquelle le choix des expressions référentielles dépend du degré de topicalité d'une entité dans le discours. Ainsi, pour référer à une expression qui sert de topique dans le discours (c'est-à-dire un référent déjà mentionné dans le contexte ou supposé connu par l'interlocuteur), on aura tendance à utiliser un

pronom plutôt qu'un nom propre. Cette hypothèse a été affinée dans la théorie de l'accessibilité d'Ariel 1990 et 1994, qui met l'accent sur les expressions anaphoriques telles que les pronoms. Selon la théorie de l'accessibilité, les formes référentielles peuvent être classées sur une échelle de complexité allant des noms propres aux expressions nulles, comme schématisé en (6) ci-dessous. Plus une entité est accessible dans le discours, moins l'expression anaphorique choisie sera complexe. À titre d'exemple, en (7a), l'antécédent *Max* est très accessible dans le discours, par conséquent il sera repris par une expression anaphorique plus réduite (*il*). En revanche, en (7b), l'antécédent *Max* est moins accessible qu'en (7a), vu la présence d'un autre antécédent potentiel (*Paul*), ce qui entraîne l'emploi d'une forme référentielle plus complexe (*ce dernier*).

(6) Echelle de l'accessibilité des référents (v. Ariel 1990 et 1994)
Référent : [+accessible] ---------- [–accessible]
Expression : pronom nul ---- pronom faible ---- pronom fort ---- nom propre

(7) a. *J'ai vu Max*$_i$. ***Il***$_i$ *semblait fatigué*. (+accessible)
b. *Paul a rencontré Max*$_i$. ***Ce dernier***$_i$ *semblait fatigué.* (–accessible)

À l'instar de l'étude de Givón 1983, Nappa et Arnold 2014 proposent deux facteurs qui seraient responsables de la proéminence dans le discours, à savoir la structure informationnelle et la fonction syntaxique. Ainsi, un référent topique sera plus accessible qu'un référent non-topique et il en va de même pour les antécédents sujet par rapport aux antécédents non-sujet. Cependant, comme la plupart des études portent sur la résolution des pronoms dans les phrases à ordre SVO (sujet-verbe-objet), on ne peut pas *a priori* bien déceler la contribution de chaque facteur. En l'absence d'études expérimentales contrôlées manipulant la structure informationnelle, on pourrait dire que les préférences du sujet nul pour un antécédent sujet préverbal peuvent être dues à la topicalité plutôt qu'à la fonction syntaxique. C'est ce qu'affirment Runner et Ibarra 2016, à partir des données sur les sujets nuls et pronominaux en espagnol, montrant que le statut de topique de l'antécédent semble jouer un rôle au-delà de sa fonction grammaticale.

Une autre théorie qui s'intéresse au choix des expressions référentielles est la théorie du centrage (*Centering Theory*, cf. Grosz *et al.* 1995, Walker et Prince 1996), dont le but est de relier la focalisation de l'attention, le choix de l'expression référentielle et la cohérence discursive (c'est-à-dire la relation sémantique qui s'établit entre les énoncés à l'intérieur d'un segment de discours : relation causale, temporelle, etc.). L'hypothèse de base est que certaines entités, appelées *centres* (ou référents discursifs dans la théorie de la représentation du discours DRT, *Discourse Representation Theory*), sont plus centrales que d'autres dans le discours. Cette propriété impose des contraintes sur l'emploi des différents types d'expressions référentielles. Ces entités sont classées selon leur saillance

discursive, qui est assimilée à la fonction syntaxique : sujet > objet direct > objet indirect > autre fonction. En lien avec la fonction syntaxique, la théorie du centrage prédit les choix des expressions référentielles en prenant en compte le type de transition : la continuité du topique ou le changement de topique. Bien que la théorie ait été proposée principalement pour les données de l'anglais, elle pourrait également expliquer les préférences relatives à l'alternance sujet nul / sujet pronominal dans les langues prodrop comme l'espagnol, l'italien, le roumain : le pronom nul serait ainsi l'expression la plus appropriée pour marquer la continuité du topique, tandis que le sujet pronominal serait plus approprié pour marquer un changement de topique. On observe ainsi dans l'exemple roumain attesté (8a) qu'on a un sujet nul dans la subordonnée temporelle, car il continue le même topique que le sujet topique de la phrase principale (*Martin Şluţ = Ø*). En revanche, en (8b), on pourrait dire que l'emploi du sujet pronominal *ea* est lié au changement de topique : l'antécédent de ce pronom (*ale unei vieţi...*) est un complément du nom dans la phrase principale et ne constitue pas le topique de celle-ci. Donc, pour marquer le passage d'un topique à l'autre, on fait appel à l'emploi du sujet pronominal.

(8) a. *Procedând în acest fel, **Martin Şluţ** şi-a încălcat promisiunea făcută anul trecut, atunci când **Ø** a fost ales în fruntea Parlamentului de la Strasburg.*
(continuité du même topique sujet)
'En agissant de cette manière, Martin Şluț a rompu la promesse qu'il avait faite l'année dernière lorsqu'il a été élu à la tête du Parlement de Strasbourg.'

b. *De exemplu, multe întrebări ale unui test numit 'Purpose in Life' sunt indicii **ale unei vieţi** realizate în mod fericit, întrucât **ea** este plină de sens.*
(changement de topique)
'Par exemple, de nombreuses questions d'un test intitulé *Purpose in Life* sont révélatrices d'une vie heureuse et épanouie, car elle est pleine de sens.'

La notion de saillance linguistique a été également mise en relation avec la résolution des pronoms pour expliquer leur référence et leur accessibilité dans le discours (Landragin 2020). Issue du domaine de la perception visuelle, la saillance suppose l'émergence d'une forme sur un fond, en particulier la mise en avant d'une entité par rapport à d'autres entités. En linguistique, la saillance permet de rendre compte du choix entre plusieurs interprétations possibles dans un contexte d'ambiguïté (en compréhension) ainsi qu'entre plusieurs expressions référentielles possibles (en production)[2].

[2] La notion de saillance est également responsable d'autres phénomènes linguistiques tels que l'ordre des compléments en français (Fahiri et Thuillier 2018), le marquage différentiel de l'objet direct en roumain (Mardale 2011), etc.

On voit donc qu'il y a plusieurs facteurs qui pourraient expliquer les contraintes préférentielles qui gèrent la résolution des expressions anaphoriques dans des contextes ambigus.

1.3. Objectifs de cet article

Dans cet article, nous voulons étudier certains de ces facteurs dans l'alternance sujet nul / sujet pronominal en roumain, vu le fait que cette langue est la seule langue romane qui ne dispose pas d'une étude quantitative exhaustive dédiée à ce thème.

La structure de l'article est la suivante : dans la section 2, nous passons en revue rapidement les études expérimentales qui ont été faites dans plusieurs langues romanes, en mettant l'accent sur la variation que l'on peut trouver au sein des langues romanes. Dans la section 3, nous présentons l'étude de corpus que nous avons menée sur le roumain, qui prend en compte plusieurs facteurs, dont certains seront discutés en détail. La section 4 comporte une discussion générale des faits discutés dans les sections précédentes, avant de conclure (section 5) sur l'importance d'une étude empirique (qui combine expérimentation et corpus) pour l'étude des préférences des locuteurs quand on a affaire à des alternances syntaxiques.

2. Études quantitatives menées sur les langues romanes

Dans cette section, nous allons passer en revue les principales études expérimentales faites sur les langues romanes, qui testent toutes l'hypothèse de la position de l'antécédent (Carminati 2002, 2005).

2.1. Variations à travers les langues

L'hypothèse de la position de l'antécédent discutée dans la section 1 a été validée par les expériences faites sur plusieurs langues romanes prodrop. Globalement, on observe un effet de partage des tâches (*division of labour*) en ce sens que, typiquement, dans une phrase SVO, les sujets nuls dans une phrase enchâssée préfèrent un antécédent sujet, tandis que les sujets pronominaux préfèrent un antécédent non-sujet.

Ainsi, Carminati 2002 montre pour l'italien une préférence très forte (76.89%) des sujets nuls dans les subordonnées temporelles quand leurs antécédents correspondent au sujet de la principale. De même, les sujets pronominaux sont fortement préférés avec un antécédent objet (82.95%). Cette distribution complémentaire prédisant l'antécédent d'un sujet anaphorique a également été mise en évidence pour le portugais européen (Fernandes *et al.* 2018) : les participants manifestent une forte préférence pour les antécédents sujet dans le cas des sujets nuls (79%) et pour les antécédents objet dans le cas des sujets pronominaux (76%).

Cependant, dans une perspective romane comparative, il semble exister une certaine variation concernant la force de l'effet de partage des tâches (v. le tableau 1 ci-dessous). La préférence pour un antécédent sujet semble être plus robuste pour les sujets nuls dans toutes les langues, malgré certaines variations (v. Filiaci *et al.* 2014 pour l'espagnol et l'italien, de la Fuente et Hemforth 2013 pour l'espagnol, Fernandes *et al.* 2018 pour le portugais européen et brésilien, Torregrosa *et al.* 2020 pour l'italien). En revanche, pour les sujets pronominaux, une plus grande variation entre les langues a été constatée (par exemple, l'espagnol est beaucoup plus flexible que l'italien pour ce qui est de l'antécédent d'un sujet pronominal : Sorace et Filiaci 2006, Filiaci *et al.* 2014, Chamorro 2018, Contemori et Di Domenico 2021, Torregrosa *et al.* 2020). Par conséquent, l'interprétation des sujets anaphoriques diffère selon les langues, en particulier dans le cas des sujets pronominaux.

	sujet nul		**sujet pronominal**	
	ant. sujet	**ant. objet**	**ant. sujet**	**ant. objet**
italien	76.89	23.11	17.05	82.95
portugais européen	79	21	24	76
espagnol	66.32	36.81	33.68	63.19
catalan	59.1	40.9	35.2	64.8
portugais brésilien	74	26	54	46

Tableau 1. Les préférences des locuteurs concernant la résolution du sujet anaphorique (en pourcentages).

Pour l'espagnol (Filiaci *et al.* 2014, de la Fuente et Hemforth 2013 *inter alia*), l'asymétrie sujet *vs* objet est assez claire : les sujets nuls préfèrent les antécédents sujet comme en italien ou en portugais européen (66.32%), tandis que les sujets pronominaux montrent une préférence pour les antécédents objet (63.19%). Les deux variantes diatopiques du portugais (européen et brésilien) présentent également une variation importante selon le partage des tâches (Fernandes *et al.* 2018). Bien qu'on remarque une préférence générale pour un antécédent sujet dans le cas du sujet nul de la subordonnée (79% en portugais européen *vs* 74% en portugais brésilien), il existe une différence significative dans l'interprétation du sujet pronominal. En portugais européen, l'objet est clairement l'antécédent préféré du sujet pronominal de la subordonnée, alors qu'en portugais brésilien il n'y a pas de préférence claire entre l'antécédent sujet et l'antécédent objet. Enfin, les préférences observées en catalan (Mayol 2010) montrent une tendance particulière, les préférences des sujets nuls pour un antécédent sujet étant légèrement plus faibles (59.1%) que celles trouvées pour les sujets pronominaux associés à des antécédents objet (64.8%).

On voit donc que, même à l'intérieur d'une même famille de langues, comme c'est le cas des langues romanes prodrop, on ne retrouve pas exactement le même comportement quant à la résolution du sujet anaphorique (sujet nul *vs* sujet pronominal), mais plutôt une microvariation, la généralisation liée à l'emploi du sujet nul étant plus forte que celle portant sur l'emploi du sujet pronominal.

2.2. Étude expérimentale sur le roumain

Comme le roumain ne disposait pas d'étude expérimentale sur l'alternance sujet nul / sujet pronominal, nous avons décidé de monter une telle expérience, qui garde le même matériel expérimental que celui testé sur l'espagnol par de la Fuente et Hemforth 2013, ainsi que par Fernandes *et al.* 2018. Ceci est particulièrement utile car nous pourrons comparer nos données à une langue avec un effet de partage des tâches faible à modéré (comme l'espagnol) ainsi qu'à une langue avec effet plus catégorique (comme le portugais européen).

Les détails de cette expérience sur le roumain sont présentés dans Istrate *et al.* 2022. Pour les besoins de cet article, nous avons décidé d'exposer tout simplement les résultats généraux qui complètent le tableau des langues romanes, comme synthétisé dans le tableau 2.

Nous pouvons observer que les tendances en roumain sont très similaires à celles trouvées en espagnol (de la Fuente et Hemforth 2013), c'est-à-dire moins robustes qu'en italien et en portugais européen avec une préférence (64.2%) des sujets nuls pour un antécédent sujet et une préférence (58.2) des sujets pronominaux pour un antécédent objet.

	sujet nul		**sujet pronominal**	
	ant. sujet	**ant. objet**	**ant. sujet**	**ant. objet**
italien	76.89	23.11	17.05	82.95
portugais européen	79	21	24	76
espagnol	66.32	36.81	33.68	63.19
catalan	59.1	40.9	35.2	64.8
roumain	64.2	35.8	41.8	58.2
portugais brésilien	74	26	54	46

Tableau 2. Les préférences (en pourcentages), incluant les résultats du roumain.

De façon générale, on observe donc que, bien que toutes ces langues soient des langues prodrop, il y a des microvariations, au-delà des généralisations qui s'y appliquent. Crucialement, ces microvariations ont pu être décelées grâce à l'expérimentation ; en l'absence des expériences psycholinguistiques, il n'aurait pas été possible de les mesurer.

3. Étude de corpus sur le roumain

Une étude empirique est complète si elle combine les données de laboratoire (c'est-à-dire l'expérimentation linguistique) avec les données attestées en corpus. Afin d'observer comment les locuteurs du roumain gèrent l'alternance sujet nul / sujet pronominal en production, dans un usage naturel de la langue, nous avons fait appel à deux corpus du roumain.

L'hypothèse qui a guidé notre étude découle des études expérimentales antérieures : malgré les ressemblances qui existent au sein d'une même famille de langues, comme la famille des langues romanes, on s'attend à ce qu'il y ait de la (micro)variation, comme on l'a observé dans la section 2.

3.1. Corpus choisis et échantillonnage

Nous avons choisi un corpus pour chaque modalité de la langue : le corpus *CoRoLa* (*The Reference Corpus of the Contemporary Romanian Language*) pour la composante orale de la langue et le corpus *Parseme-ro 1.2* pour la composante écrite.

Les textes oraux de *CoRoLa* que nous avons retenus pour notre étude sont principalement des enregistrements professionnels provenant de différentes sources (actualités et conversations radio, interviews, contes de fées racontés oralement), qui disposent d'une transcription écrite et qui englobent 152 heures d'enregistrement.

Le corpus *Parseme-ro 1.2* est un corpus écrit journalistique homogène, qui regroupe des textes tirés d'*Agenda newspaper* et dont la taille est de 56.703 phrases et 1.015.624 mots.

L'échantillonnage a été fait aléatoirement pour chaque modalité. Comme le corpus *Parseme-ro* est annoté syntaxiquement, nous avons fait appel à des formules de requête pour le recueil automatique des données. En revanche, comme le corpus *CoRoLa* ne dispose pas d'une annotation syntaxique, nous avons été contraints d'utiliser un autre moyen d'extraction des données : nous avons regardé la fréquence des verbes, en vérifiant quels sont les verbes les plus courants en roumain et nous avons retenu seulement les occurrences contenant des sujets pronominaux et nuls (éliminant les phrases à sujets lexicaux).

Notre échantillonnage tient compte non seulement du type de sujet (sujet nul *vs* sujet pronominal), mais aussi du type de phrase (phrase simple *vs* phrase complexe, qui contient une subordonnée). Nous avons ainsi deux échantillons : un premier échantillon de 400 phrases simples (200 phrases à sujet nul et 200 phrases à sujet pronominal, avec une distribution équilibrée selon la modalité : 100 occurrences pour le corpus oral et 100 occurrences pour le corpus écrit) et un deuxième échantillon de 368 phrases complexes (200 phrases à sujet nul et 168 phrases à sujet pronominal).

Le roumain est une langue à ordre des mots relativement libre (*GALR* 2005, Pană-Dindelegan 2003, 2013), donc le sujet peut apparaître tant en position préverbale qu'en position postverbale. Pour que notre étude puisse être comparée

aux autres études faites en corpus, l'extraction des données a privilégié les sujets en position préverbale, et plus particulièrement les sujets préverbaux des phrases déclaratives. Tous les sujets pris en compte sont référentiels, ce qui veut dire que nous avons éliminé les autres occurrences de sujet nul, par exemple les sujets nuls des verbes météorologiques.

Quelques exemples de nos échantillons figurent en (9) pour les phrases simples et en (10) pour les phrases complexes.

(9) a. *Deci, **Ø** ar intra direct pe teritoriul unui stat NATO*. (*CoRoLa*)
'Il pénétrerait donc directement sur le territoire d'un État membre de l'OTAN.'
b. ***Ø** Repar TV color şi alb-negru la domiciliul clientului.* (*Parseme-ro*)
'Je répare des télés couleur et noir et blanc au domicile du client.'
c. ***Eu** am lucrat mulţi ani la serviciul de lansare*. (*CoRoLa*).
'J'ai travaillé pendant de nombreuses années dans le service de lancement.'
d. ***El** nu va putea moşteni şi pe rudele sale fireşti*. (*Parseme-ro*)
'Il ne pourra pas hériter de ses parents naturels.'

(10) a. *N-aş vrea să exagerăm puţin cu responsablitatea individuală, pentru că **Ø** a fost invocată foarte des*. (*CoRoLa*)
'Je ne voudrais pas qu'on exagère un peu avec la responsabilité individuelle, parce qu'elle a été invoquée très souvent.'
b. *Nimic nu mai contează de vreme ce **el** o cere în căsătorie*. (*Parseme-ro*)
'Rien d'autre n'a d'importance puisqu'il la demande en mariage.'

3.2. Facteurs pris en compte pour l'annotation (en lien avec les autres études de corpus)

Les facteurs pris en compte pour l'annotation concernent les propriétés du verbe, les propriétés du sujet et celles de la phrase en entier. La liste de ces facteurs est donnée dans le tableau 3 ci-dessous.

Niveau	**Facteurs pris en compte**
pour la phrase	contexte gauche, polarité, sujet nul précédent
pour la subordonnée	position (finale/initiale), type de subordonnée (circonstancielle, complétive, relative)
pour la subordonnée circonstancielle	connecteur, relation discursive
pour le verbe	lemme, agentivité, voix, mode, temps, personne
pour le sujet	nombre, genre, animéité, type (nul/pronom), place du sujet (principale/subordonnée)
pour l'antécédent du sujet	un/plusieurs, type d'antécédent (nul/pronom/nom), fonction syntaxique

Tableau 3. Les facteurs pris en compte pour l'annotation.

On a donc une analyse multifactorielle, qui prend en compte plusieurs facteurs à la fois, ce qui fait la différence par rapport aux études expérimentales

présentées précédemment, qui prenaient en compte essentiellement deux facteurs : la fonction syntaxique de l'antécédent (sujet *vs* non-sujet) et le type de réalisation du sujet (nul *vs* pronominal), v. l'hypothèse de la position de l'antécédent de Carminati 2002.

Les différents facteurs listés dans le tableau 3 sont pris en compte dans les autres études de corpus dédiées aux réalisations du sujet à travers les langues. Pour les buts de cet article, on va insister ici uniquement sur les facteurs qui seront analysés dans la section Résultats.

En particulier, il semble que la personne grammaticale et l'animéité ont une forte influence sur l'alternance sujet nul *vs* sujet pronominal, comme le montrent d'autres études de corpus (Duarte 2000, Manjón Cabeza-Cruz *et al.* 2016, Correa Soares *et al.* 2020, Ávila et Segura Lores 2022). En particulier, en espagnol oral (v. Manjón Cabeza-Cruz *et al.* 2016 pour l'espagnol parlé de Grenade et Ávila et Segura Lores 2022 pour l'espagnol parlé de Malaga), les études de corpus montrent une fréquence très forte de la première personne du singulier avec les sujets pronominaux. Avec des résultats proches de l'espagnol, Duarte 2000 et Correa Soares *et al.* 2020 montrent qu'en portugais brésilien parlé et en portugais européen écrit il y a une fréquence élevée des sujets pronominaux aux personnes du discours. En revanche, en portugais brésilien, les sujets nuls sont produits plus souvent à la troisième personne, avec un antécédent [-animé] et [-spécifique].

Dans ce qui suit, nous allons étudier de plus près les données du roumain, pour voir s'il suit les mêmes tendances que celles observées dans les autres langues romanes. S'il suit les mêmes tendances, on s'attend donc à ce que les sujets nuls soient plus fréquents à la 3^{e} personne (surtout avec des référents inanimés au singulier) et que les sujets pronominaux soient plus fréquents aux personnes du discours.

3.3. Résultats

Dans cette section, nous allons insister sur les effets les plus intéressants, qu'on n'a pas pu observer dans les études expérimentales que nous avons menées sur le roumain jusqu'à maintenant.

Un des effets remarquables est celui de la personne du sujet, comme on peut le voir dans la figure 1. Généralement, les personnes du discours (c'est-à-dire la première et la deuxième personne) ont été plus fréquentes que la troisième personne. Si l'on prend en considération la variable type de sujet (sujet nul *vs* sujet pronominal), on observe que le sujet nul est plus fréquent avec les personnes du discours (72.6%), alors que le sujet pronominal est majoritairement associé à la troisième personne (80.6%). Pour l'analyse statistique des résultats, nous avons fait appel au modèle linéaire mixte de régression logistique (Bates *et al.* 2015) en R (R Development Core Team 2015). Les statistiques inférentielles nous montrent elles aussi un effet significatif de la personne ($p < .001$), confirmant ainsi les résultats des statistiques descriptives que l'on peut voir dans la figure 1.

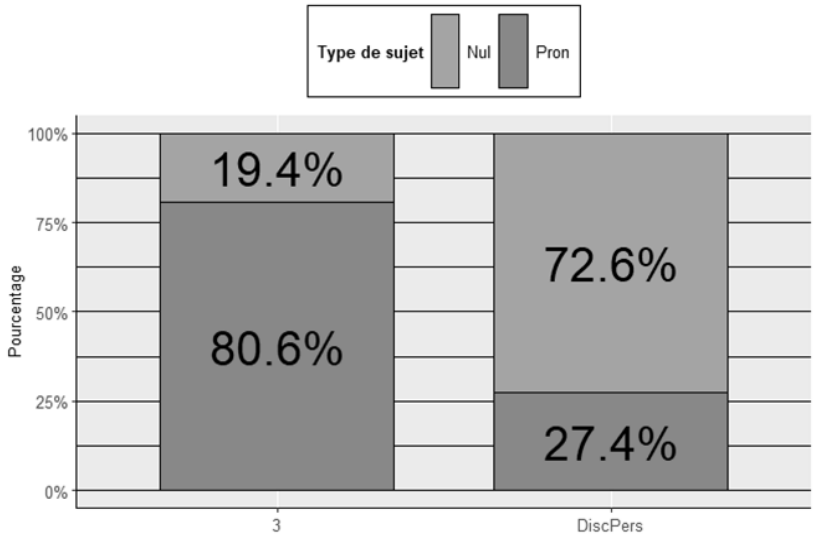

Figure 1. L'effet de la personne dans les corpus roumains.

La distribution asymétrique du sujet en fonction de la personne dépend également de la modalité (écrite ou orale), comme on peut le voir dans les figures 2 et 3. À l'écrit (figure 2), l'effet de personne est plus robuste, avec une fréquence plus élevée des sujets nuls (91.1%) aux personnes du discours et des sujets pronominaux à la troisième personne (91.9%). La même tendance est observée également à l'oral (figure 3), bien que moins catégorique qu'à l'écrit : 58% sujets nuls aux personnes du discours et 65.2% sujets pronominaux à la troisième personne. La régression logistique montre une interaction très forte ($p < .001$) entre la personne du sujet et la modalité (type de corpus : écrit *vs* oral).

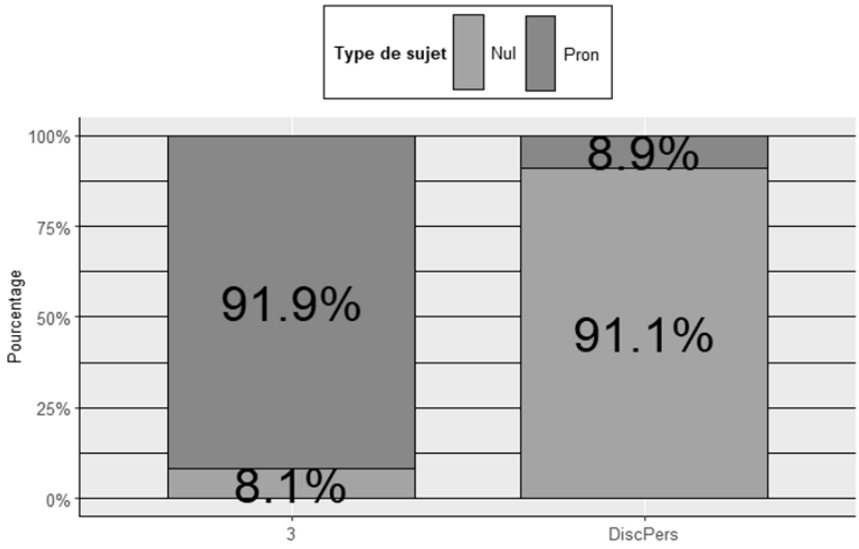

Figure 2. L'effet de la personne dans le corpus écrit.

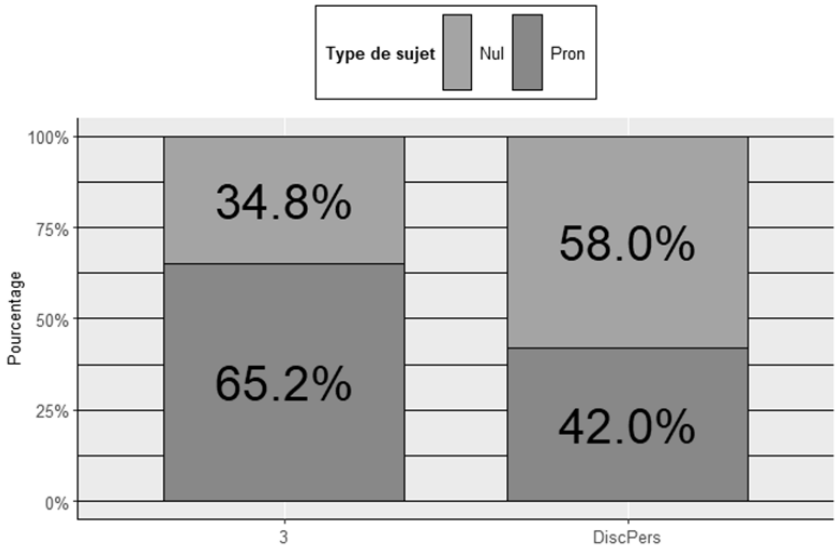

Figure 3. L'effet de la personne dans le corpus oral.

En plus de la personne du sujet et de la modalité, un autre fait intéressant est lié à l'animéité. Comme on peut le voir dans la figure 4, nous observons une fréquence plus élevée des sujets pronominaux qui ont le trait [-animé] : 78%. Dans le cas des référents animés, les résultats ne montrent pas de différence dans la distribution des sujets nuls et pronominaux.

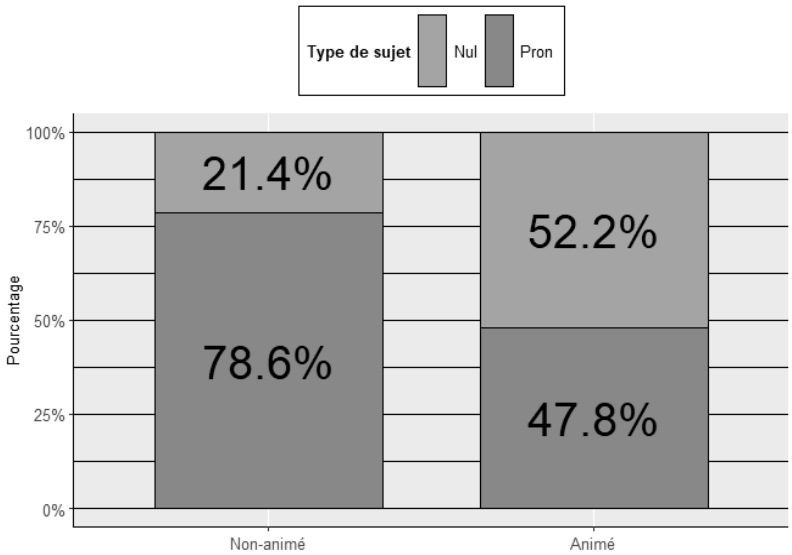

Figure 4. L'effet de l'animéité dans les corpus roumains.

4. Discussion générale

L'existence d'un effet significatif de la personne en roumain quant à l'alternance sujet nul / sujet pronominal pourrait être expliquée en termes d'accessibilité référentielle ou de saillance (Ariel 1990, 1994). Comme les personnes du discours représentent intrinsèquement des référents humains, ils sont plus saillants dans le discours et, par conséquent, seront repris par des expressions moins complexes telles que les sujets nuls. En revanche, la 3^{e} personne pourrait renvoyer également à des antécédents [-animés], [-humains] ou [-spécifiques], moins saillants dans ces contextes, ce qui pourrait expliquer pourquoi ils sont très souvent repris par des sujets pronominaux, c'est-à-dire par des expressions plus complexes. En explorant l'interaction entre la personne et le type de sujet en fonction de la modalité (écrite / orale), nous avons montré que la fréquence générale s'applique à l'écrit ainsi qu'à l'oral, mais elle est beaucoup plus catégorique en roumain écrit.

Néanmoins, il n'est pas suffisamment clair pourquoi les tendances observées en roumain s'opposent à celles trouvées dans d'autres langues romanes telles que le portugais brésilien parlé (Correa Soares *et al.* 2020), le portugais européen écrit (Duarte 2000) ou l'espagnol parlé de Malaga et de Grenade (Manjón Cabeza-Cruz *et al.* 2016 et Ávila et Segura Lores 2022). En essayant de rendre compte de la distribution asymétrique des sujets selon la personne en portugais brésilien parlé, Correa Soares *et al.* (2020) remettent en question l'hypothèse de la hiérarchie référentielle prise en compte par d'autres études de corpus antérieures (Cyrino *et al.* 2000, Kato et Negrão 2000). Contrairement à la théorie de l'accessibilité (Ariel 1990, 1994), cette hypothèse postule que plus une entité est référentielle, plus elle sera reprise par des formes complexes, c'est-à-dire par des sujets pronominaux. Des études de corpus parallèles et des expériences plus contrôlées seraient très utiles pour pouvoir affiner les différences entre les langues et trancher en faveur de l'une ou de l'autre des possibilités dont on dispose pour expliquer ces effets.

On voit ainsi que les études de corpus, tout comme les études expérimentales, montrent qu'un même phénomène syntaxique (à savoir l'alternance sujet nul / sujet pronominal) donne lieu à des généralisations qui s'appliquent à plusieurs langues, mais aussi à des (micro)variations, qui sont spécifiques à une langue ou un sous-groupe de langues.

Cette étude de corpus sur le roumain n'est qu'une petite partie d'une étude plus approfondie, qui prend en compte tous les facteurs mentionnés dans la section 3.2., mais elle est suffisante pour démontrer l'importance d'un tel type d'étude pour mesurer de petits effets qu'on ne verrait pas autrement.

5. Conclusions

Cet article montre l'importance d'une approche empirique pour l'étude des alternances syntaxiques, comme l'alternance sujet nul / sujet pronominal en roumain, qui ne se laissent pas décrire par des contraintes syntaxiques binaires, catégoriques (v. la distinction classique *grammatical* vs *agrammatical*), mais plutôt par des contraintes non-catégoriques, qui rendent compte des préférences des locuteurs pour une structure ou l'autre. Ces contraintes préférentielles peuvent être mesurées statistiquement grâce aux expériences psycholinguistiques qui suivent un protocole expérimental très strict et/ou grâce aux études de corpus, qui permettent des analyses quantitatives très poussées.

Concernant l'alternance sujet nul / sujet pronominal, cet article montre que les études de corpus tout comme les travaux expérimentaux nous permettent d'observer des (micro)variations au sein des langues romanes prodrop. Si l'hypothèse de la position de l'antécédent (Carminati 2002, 2005) semble expliquer (au moins en partie) les préférences des locuteurs (à savoir, préférence du sujet nul pour un antécédent sujet *vs* préférence du sujet pronominal pour un antécédent non-sujet), elle doit être complétée par l'appel à d'autres facteurs. Un de ces facteurs est celui de la personne, comme on a pu le voir grâce à l'investigation du phénomène en corpus.

L'alternance sujet nul / sujet pronominal n'est donc pas une étude simple à faire, car elle doit inclure une analyse multifactorielle, qui fait appel à différents niveaux linguistiques (syntaxiques, discursifs, morphologiques, etc.).

Références bibliographiques

Ariel, Mira (1990) – *Accessing Noun-Phrase Antecedents*, Vol. 96, Londres, Routlege.

Ariel, Mira (1994) – "Interpreting anaphoric expressions: A cognitive versus a pragmatic approach", *Journal of linguistics*, n° 30(1), p. 3-42.

Arnold, Jennifer E., Eisenband, Janet G., Brown-Schmidt, Sarah, & Trueswell, John. (2000) – "The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking", *Cognition*, n° 76(1), B13-B26.

Ávila, Antonio Manuel, Segura Lores, Alba. (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*, n° 10(2), p. 57-94.

Bates, Douglas, Mächler, Martin, Bolker Ben, Walker, Steve (2015) – "Fitting Linear Mixed-Effects Models Using lme4", *Journal of Statistical Software*, n° 67(1), p. 1-48.

Bîlbîie, Gabriela, Faghiri, Pegah, Thuilier, Juliette (2021) – « Syntaxe quantitative et expérimentale : objets et méthodes », *Langages*, n° 3, p. 7-24.

Bresnan Joan & Ford Marilyn (2010) – "Predicting syntax : processing dative constructions in American and Australian varieties of English", *Language*, n° 86(1), p. 168-213.

Carminati, Maria Nella (2002) – *The processing of Italian subject pronouns*, University of Massachusetts Amherst, Phd dissertation.

Carminati, Maria Nella (2005) – "Processing reflexes of the Feature Hierarchy (Person> Number> Gender) and implications for linguistic theory", *Lingua*, Vol. 115, n° 3, p. 259-285.

Chamorro, Gloria (2018) – "Offline interpretation of subject pronouns by native speakers of Spanish", *Glossa: a journal of general linguistics*, n° 3(1), p. 1-16.

Contemori, Carla, Di Domenico, Elisa (2021) – "Microvariation in the division of labor between null and overt-subject pronouns: the case of Italian and Spanish", *Applied Psycholinguistics*, n° 42(4), p. 997-1028.

Correa Soares, Eduardo, Miller, Philip, Hemforth, Barbara (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*, n° 36, document consulté en ligne : https://www.scielo.br/j/delta/a/w4g4SLvCgwbvmwQvckx5hxL/?lang=en.

Correa Soares, Eduardo (2017) – *Anaphors in discourse: anaphoric subjects in Brazilian Portuguese*, Thèse de doctorat, Université Sorbonne Paris Cité.

Cyrino, Sonia Maria Lazzarini, Duarte, Maria Eugênia Lamoglia, Kato, Mary Aizawa (2000) – "Visible subjects and invisible clitics in Brazilian Portuguese", *in* Kato, M. A. & Negrão, E. V. (eds.), *Brazilian Portuguese and the Null Subject Parameter*, Frankfurt am Main, Vervuert/ Madrid, p. 55-73.

De la Fuente, Israel, and Barbara Hemforth (2013) – "Effects of clefting and left-dislocation on subject and object pronoun resolution in Spanish", Jennifer Cabrelli Amaro *et al*. (ed.), *Selected proceedings of the 16th Hispanic linguistics symposium*, Somerville, p. 27-45.

Dobrovie-Sorin, Carmen, Giurgea, Ion (eds.) (2013) – *A reference grammar of Romanian: Volume 1: The noun phrase*, Vol. 207, Amsterdam, John Benjamins Publishing.

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*, Vervuert Verlagsgesellschaft, p. 17-36.

Faghiri, Pegah, Thuilier, Juliette (2018) – « Ordre des compléments postverbaux en français : poids et accessibilité discursive », *SHS web of conferences*, EDP Sciences, Vol. 46, document consulté en ligne : https://www.shs-conferences.org/articles/shsconf/pdf/2018/07/shsconf_cmlf2018_14008.pdf.

Fernandes, Eunice G., Luegi, Paula, Correa Soares, Eduardo, de la Fuente, Israel, Hemforth, Barbara (2018) – "Evidence from Brazilian and European Portuguese", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 44, n° 12, p. 1986-2008.

Filiaci, Francesca, Sorace, Antonella, Carreiras, Manuel (2014) – "Anaphoric biases of null and overt subjects in Italian and Spanish: a cross-linguistic comparison", *Language, Cognition and Neuroscience*, Vol. 29, n° 7, p. 825-843.

Givón, Talmy (1983) – "Topic continuity in discourse: The functional domain of switch reference", *Switch reference and universal grammar*, John Haiman & Pamela Munro (eds.), Philadelphia, John Benjamins, p. 51-82.

Goldberg, Adele E. (1995) – *Constructions: A Construction Grammar Approach to Argument Structure*, The University of Chicago Press, Chicago and London.

Grosz, Barbara J., Joshi, Aravind K., Weinstein, Scott (1995) – "Centering: A framework for modelling the local coherence of discourse", *Computational linguistics*, n° 21(2), p. 203-225.

Guţu-Romalo, Valeria (coord.) (2005) – *Gramatica limbii române*, București, Editura Academiei Române.

Holler, Anke, Suckow, Katja (2016) – *Empirical Perspectives on Anaphora Resolution*, Berlin, Boston, De Gruyter.

Istrate, Fabian, Abeillé, Anne, & Hemforth, Barbara (2022) – "The Position of Antecedent Hypothesis in Romanian subject alternation: two experiments", *Going Romance*, Barcelone, texte en ligne : https://cnrs.hal.science/hal-03924721/document.

Kato, Mary Aizawa, Negrão, Esmeralda Vailati (2000) – *Brazilian Portuguese and the Null Subject Parameter*, Frankfurt am Main, Vervuert/ Madrid.

Kehler, Andrew, Rohde, Hannah (2019) – "Prominence and coherence in a Bayesian theory of pronoun interpretation", *Journal of Pragmatics*, n° 154, p. 63-78.

Keller, Frank (2000) – *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality* (Doctoral dissertation), University of Edinburgh.

Landragin, Frédéric (2020) – « La saillance : origines perceptives, applications linguistiques, enjeux interdisciplinaires », *Semen*, n° 49, document consulté en ligne : https://journals.openedition.org/semen/14430.

Manjón-Cabeza Cruz, Antonio, Pose Furest, Francisca, & Sánchez García, Francesco José (2016) – "Factores determinantes en la expresión del sujeto pronominal en el corpus PRESEEA de Granada", *Boletín de filología*, n° 51(2), p. 181-207.

Mardale, Alexandru (2011) – Aspects de la saillance linguistique en roumain (1), *Analele Universităţii din Bucureşti*, n° 60, p. 67-83.

Mayol, Laia (2010) – "Contrastive pronouns in null-subject Romance languages", *Lingua*, Vol. 120, n° 10, p. 2497-2514.

Nappa, Rebecca, & Arnold, Jennifer E. (2014) – "The road to understanding is paved with the speaker's intentions: Cues to the speaker's attention and intentions affect pronoun comprehension", *Cognitive Psychology*, n° 70, p. 58-81.

Pană Dindelegan, Gabriela (2003) – Elemente de gramatică. Dificultăţi, controverse şi noi interpretări, București, Humanitas.

Pană Dindelegan, Gabriela, Maiden, Martin (ed.) (2013) – *The grammar of Romanian*, Oxford University Press.

Runner, Jeffrey T., Ibarra, Alyssa (2016) – "Information structure effects on null and overt subject comprehension in Spanish", *Empirical perspectives on anaphora resolution*, p. 87-112.

Sorace, Antonella, Francesca, Filiaci (2006) – "Anaphora resolution in near-native speakers of Italian", *Second language research*, n° 22(3), p. 339-368.

Thuilier, Juliette (2012) – *Contraintes préférentielles et ordre des mots en français*, Thèse de doctorat, Université Paris-Diderot-Paris VII.

Torregrossa, Jacopo, Maria Andreou, Bongartz, Christiane M. (2020) – "Variation in the use and interpretation of null subjects: A view from Greek and Italian", *Glossa: a journal of general linguistics*, Vol. 5, n° 1, p. 1-28.

Walker, Marilyn A., Prince, Ellen F. (1996) – "A bilateral approach to givenness: A hearer-status algorithm and a centering algorithm", in Thorstein Freitheim & Jeanette K. Gundel (eds.), *Reference and referent accessibility*, Amsterdan, John Benjamins, p. 291-306.