Linguistic Markers of Catalan Substitution

Jana Brusés

Agenda

1. Introduction
2. Background
3. Significance
4. Research Question
5. Theoretical Perspective
6. Data
7. Analysis
8. References

1 Introduction

- Goal 1: Examine morphological, syntactic, lexical, semantic and text-level changes
- Goal 2: Quantify language shift and determine key symptoms of Catalan’s endangerment.
- Find empirical evidence in language of its substitution

2 Background

- Use has dropped:

  43,1% (2007) → 25,1% (2022)

- Habitual language
  of less than 1/3 of the population

Territories where Catalan/Valencian is spoken and is official

Territories where Catalan/Valencian is spoken but is not official

Territories where Catalan/Valencian is not historically spoken but is official

Standard Catalan is classified as Vulnerable by the UNESCO Atlas of the World's Languages in Danger [3]

2 Background

- **Before Spanish Civil war and the dictatorship** that followed, Catalan was the **predominant** language in every domain.
- **During the dictatorship** it was **forbidden**, persecuted and eliminated from any public perspective. Spanish was imposed as only language.
- **After the dictatorship**, use was picked up, but since then, **Catalan is perceived as a subordinate** language.

3 Significance

- Catalan is vulnerable at the least endangered level.
- Catalan is not at a sudden death position, but **it is sick** and we need to find treatment and solutions.

More than just about Catalan:

“Something that all endangered languages have in common is that nobody is aware of it until it’s too late.**”** (Junyent, 2023)

Being warned gives us the chance to **look for markers and signs** of language substitution and endangerment.

4 Research Question

- Can we quantify and evaluate Language Substitution through Linguistic Markers?
- Can Linguistic Changes at different levels work as indicators of Language Substitution?

5 Theoretical Perspective

The Empirical Indicators:

- Misconception: word incorporations are the most remarkable marker of a language substitution or endangerment
- Junyent proposes the following symptoms to diagnose Catalan as a tongue threatened by language shift.

Language-threatening losses

1. Loss of word classes absent in the dominant language.
2. Time and space lexicon modification.
3. Syntactic restructuring.

6 Data

Choices and considerations:

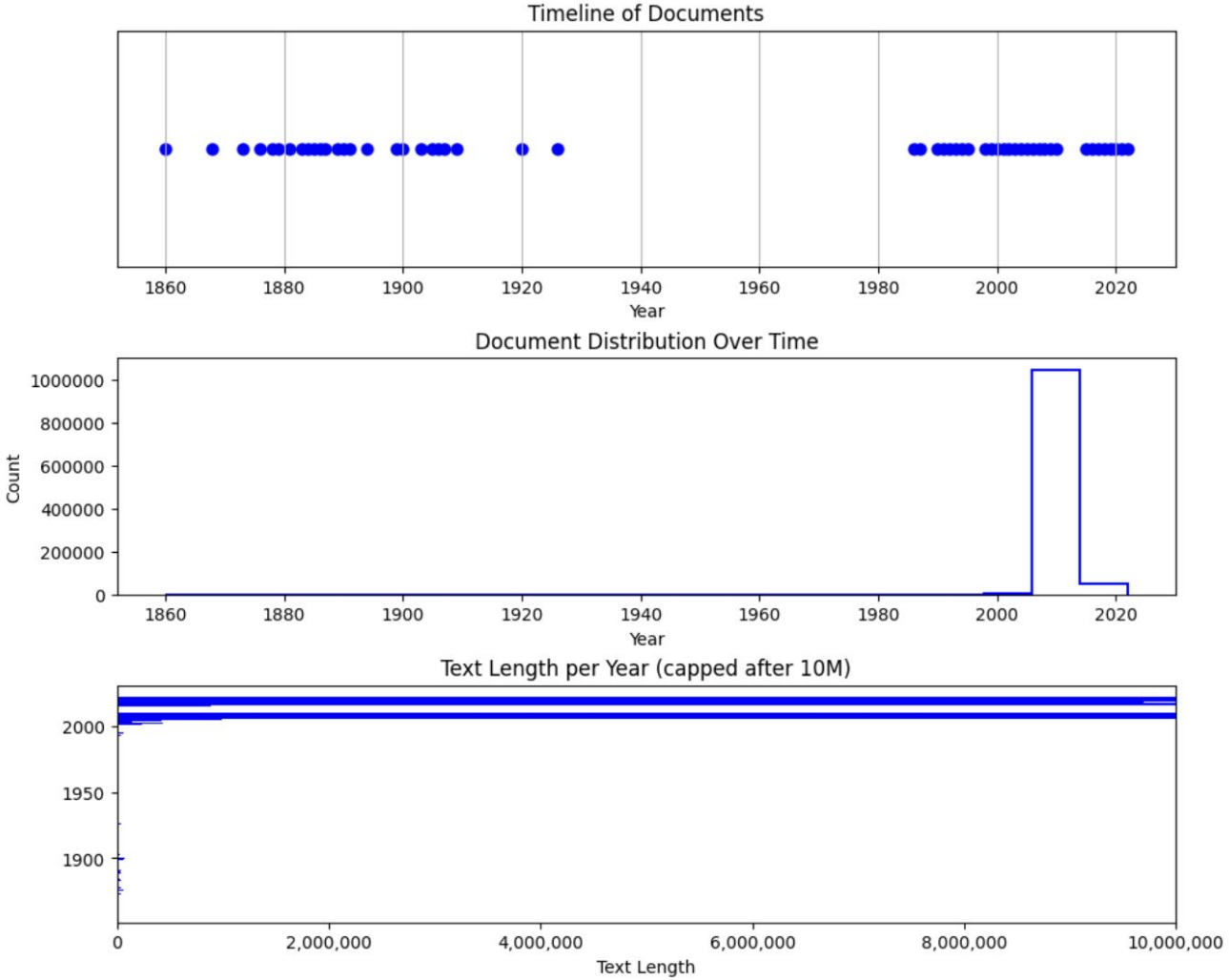
- Written language always represents a later stage than the actual appearance of any linguistic phenomenon
- Spoken would be best but is not available.
- → Solution: **Written records derived from spoken works.**

6 Data

Corpora Overview:

Corpus	Years	Notes
CTIC	1832–1926	28 Ceremonial speeches
Parlament Parla	2007–2018	700 hours total Parliament sessions
ParlaMint-ES-CT	2015–2022	Parliament corpus Multilingual/Comparable
Radioteca.cat	1985–2025	Radio broadcast transcripts

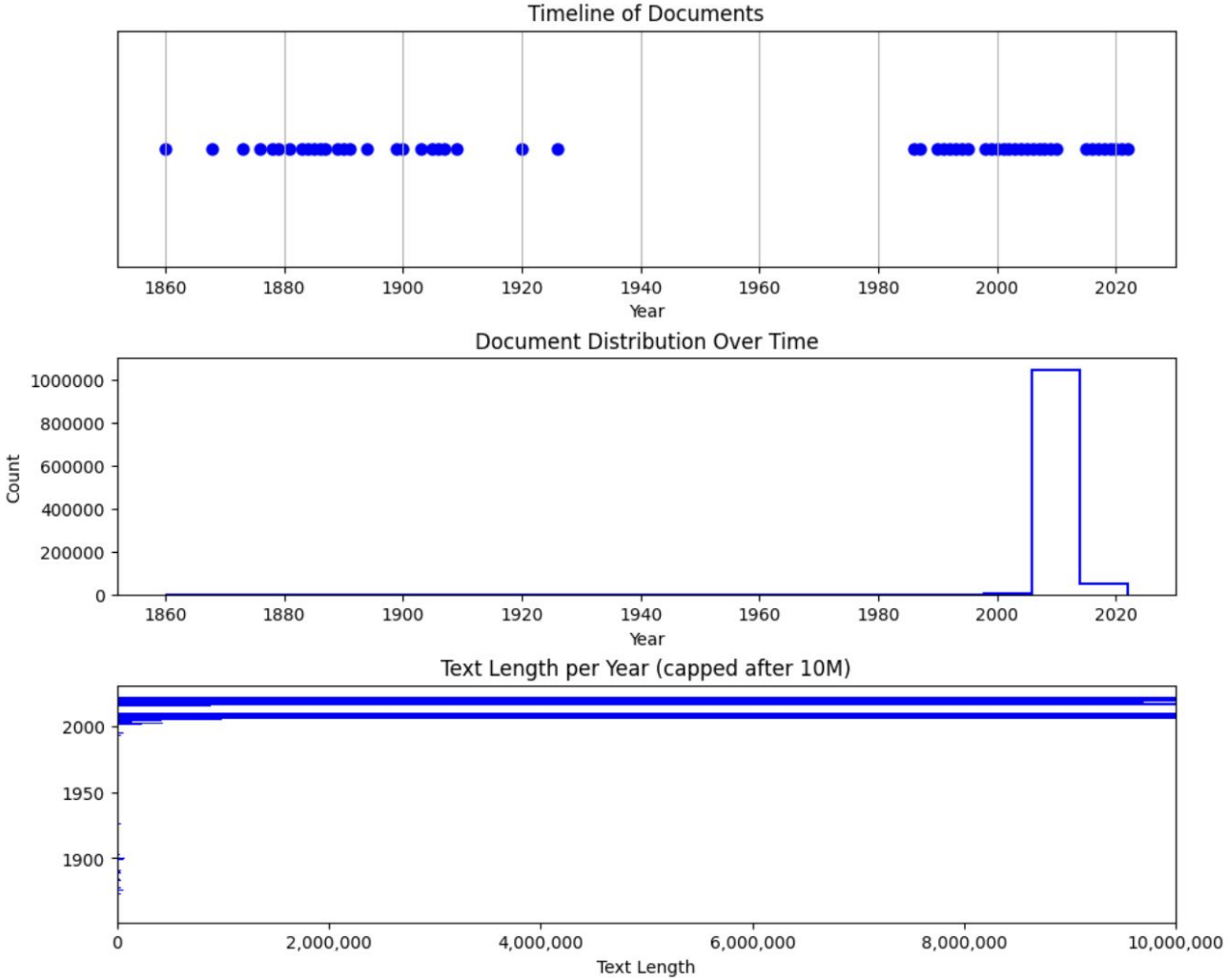
* Included to bridge the gap from 1985 to 2007

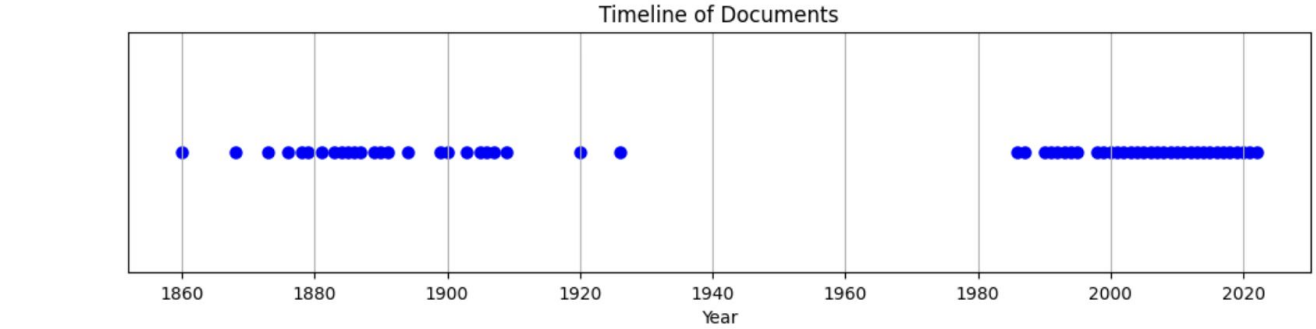

Timeline of Documents
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year
Document Distribution Over Time
Count
1000000
800000
600000
400000
200000
0
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year
Text Length per Year (capped after 10M)
2000
1950
1900
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
Text Length

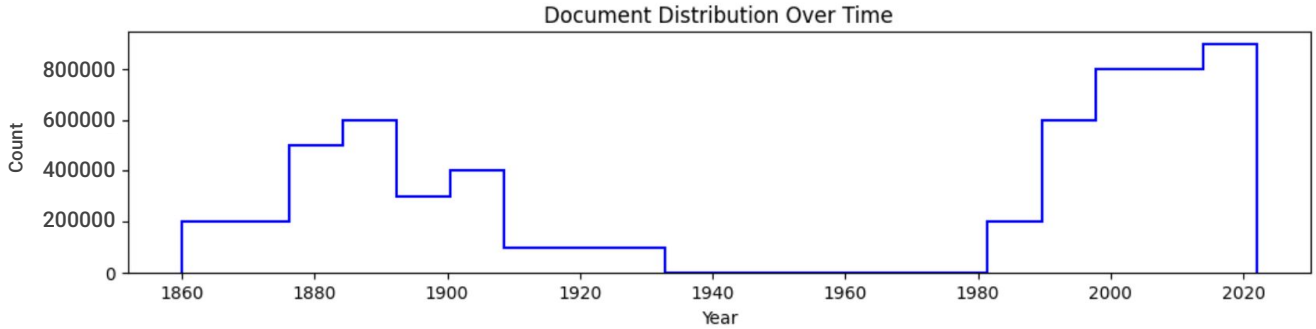
6 Data

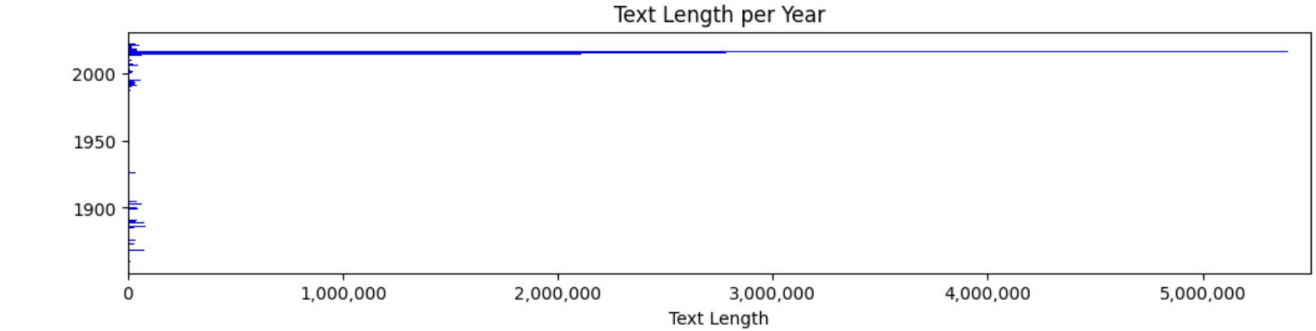
Resolving imbalance while avoiding biases:

- → Solution 1: **Limiting Individual Contribution** to 1500 characters per individual contributor to avoid speaker-specific information
- → Solution 2: **Balancing Program Contribution**. Limited Radioteca contributions to 1500 characters per episode for less topic-specific data.


Timeline of Documents
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year
Document Distribution Over Time
Count
1000000
800000
600000
400000
200000
0
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year
Text Length per Year (capped after 10M)
2000
1950
1900
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
Text Length


Timeline of Documents
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year


Document Distribution Over Time
Count
800000
600000
400000
200000
0
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year


Text Length per Year
2000
1950
1900
0
1,000,000
2,000,000
3,000,000
4,000,000
5,000,000
Text Length

6 Data

Resulting complete data:

- 3 billion tokens
- 75 thousand text-pieces
  Single speaker contribution

	Year	Text_len	Len_toks
count	75480.000000	75480.000000	75480.000000
mean	2008.832658	187.739043	40.592780
std	3.064608	611.358140	122.724208
min	1860.000000	0.000000	0.000000
25%	2008.000000	38.000000	9.000000
50%	2009.000000	95.000000	21.000000
75%	2010.000000	222.000000	48.000000
max	2022.000000	73881.000000	14727.000000

7 Analysis - Dealing with uneven distribution

- Difference in data sizes for each year makes row counts not very valuable.
- → Solution: Counts are evaluated considering the size of the amount of tokens of the year they belong to.
- Counts are turned into proportion percentage out of overall data amount for the respective year.

7 Analysis

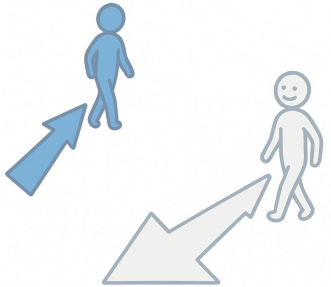
1. Loss of Word Classes
2. Time and Space Lexicon Modification
3. Syntactic restructuring
4. Text-level changes

Today's focus:

2. Time and Space Lexicon Modification
→ **Loss of directional distinction (*anar* vs. *venir*)**

3. Syntactic restructuring
→ **Verb pronominalization**

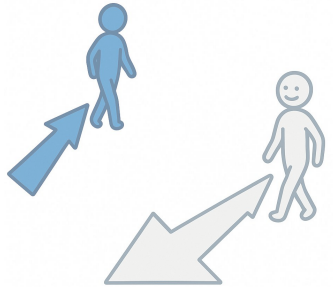
7 Analysis - Anar vs. Venir



Theoretical motivation:

- Catalan distinguishes **motion away** (*anar*) vs. **motion toward** (*venir*) the speaker.
- In Spanish the contrast is less strict.
- If one verb increases where both were once used, it may indicate **semantic erosion or substitution**.

7 Analysis - Anar vs. Venir



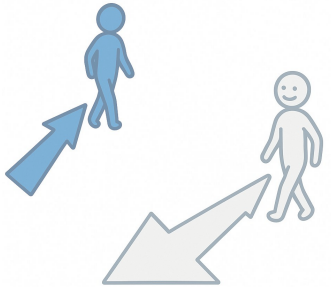
Hypothesis:

Substitution would manifest with one of the two verbs expanding its use into contexts formerly reserved for the other.

Method:

- Verbs are highly inflected (tense, person, number).
- → Solution: **Lemmatization** through Stanza's Catalan model.
- This lets us find all forms of *anar* and *venir* via their lemma.

7 Analysis - Anar vs. Venir

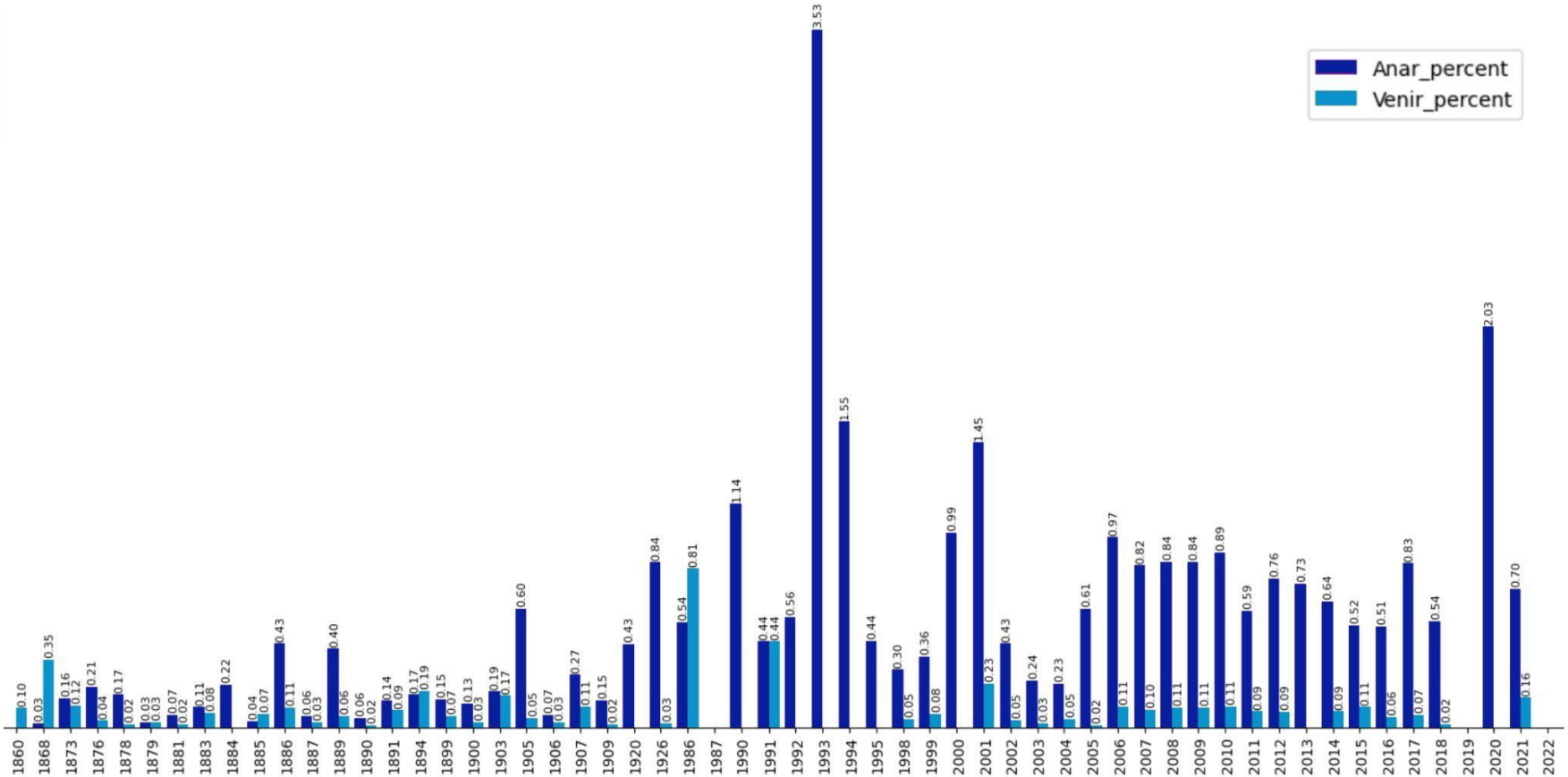


Statistical Results

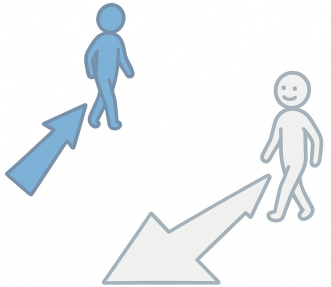
Anar shows a **significant positive correlation** with time
ρ = 0.5659, p-value = 0.0000

Venir shows **no significant change**
ρ = –0.0886, p-value = 0.5082

The **difference** between *anar* – *venir* also increases significantly
ρ = 0.5594, p-value = 0.0000


Anar_percent
Venir_percent
1860 1868 1873 1876 1878 1879 1881 1883 1884 1885 1886 1887 1889 1890 1891 1894 1899 1900 1903 1905 1906 1907 1909 1920 1926 1986 1987 1990 1991 1992 1993 1994 1995 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022

7 Analysis - Anar vs. Venir



Interpretation

- *Anar* is increasing in frequency over time
- *Venir* is stable

- → *Anar* may be **broadening in use**, not replacing *Venir*
- This supports an **extension** of *anar*

7 Analysis - Verb Pronominalization

Theoretical motivation:

- Reflexive/pronominal verbs in Catalan attach clitics (*pronoms febles*) “unstressed pronouns”

- *“Em vesteixo” = “I get dressed myself”*

- Spanish influence may trigger **over-pronominalization**,

  e.g.: *“puja’t a la bici”* instead of *“puja a la bici”*

7 Analysis - Verb Pronominalization

Hypothesis:

Due to Spanish influence Verb Pronominalization might be increasing. We might be pronominalizing movement verbs that are pronominal in Catalan's history.

7 Analysis - Verb Pronominalization

Method:

- Tokenization and POS tagging

Proclitics (separate token)	
Sentence	1 tokens =====
)	text: T’
)	text: ho
)	text: donaré
)	text: demà
)	text: .

Enclitics (same token)	
Sentence	2 tokens =====
)	text: me’n
)	text: poses
)	text: una
)	text: ?

7 Analysis - Verb Pronominalization

Method:

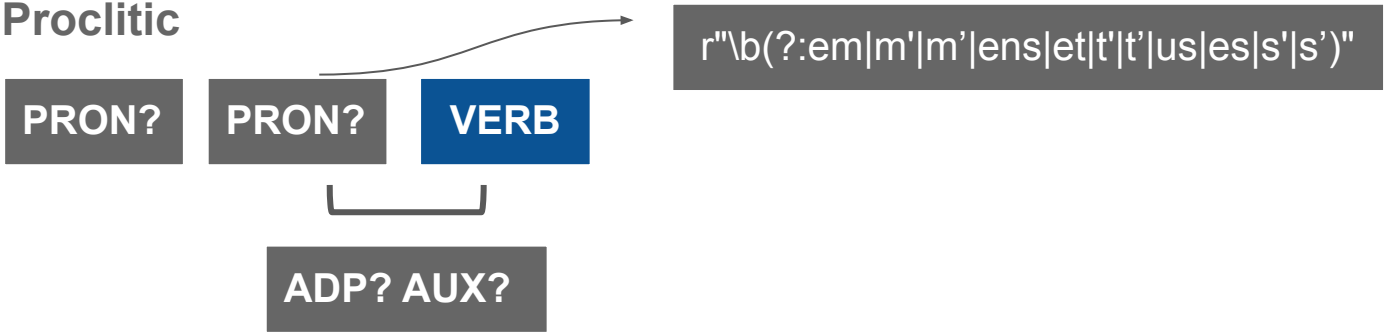
1. Tokenize and POS tag using stanza
2. Create a list of tuples containing (word, POS tag) pairs
3. Apply pronominal verbs finding function
4. Return pronominal verbs list

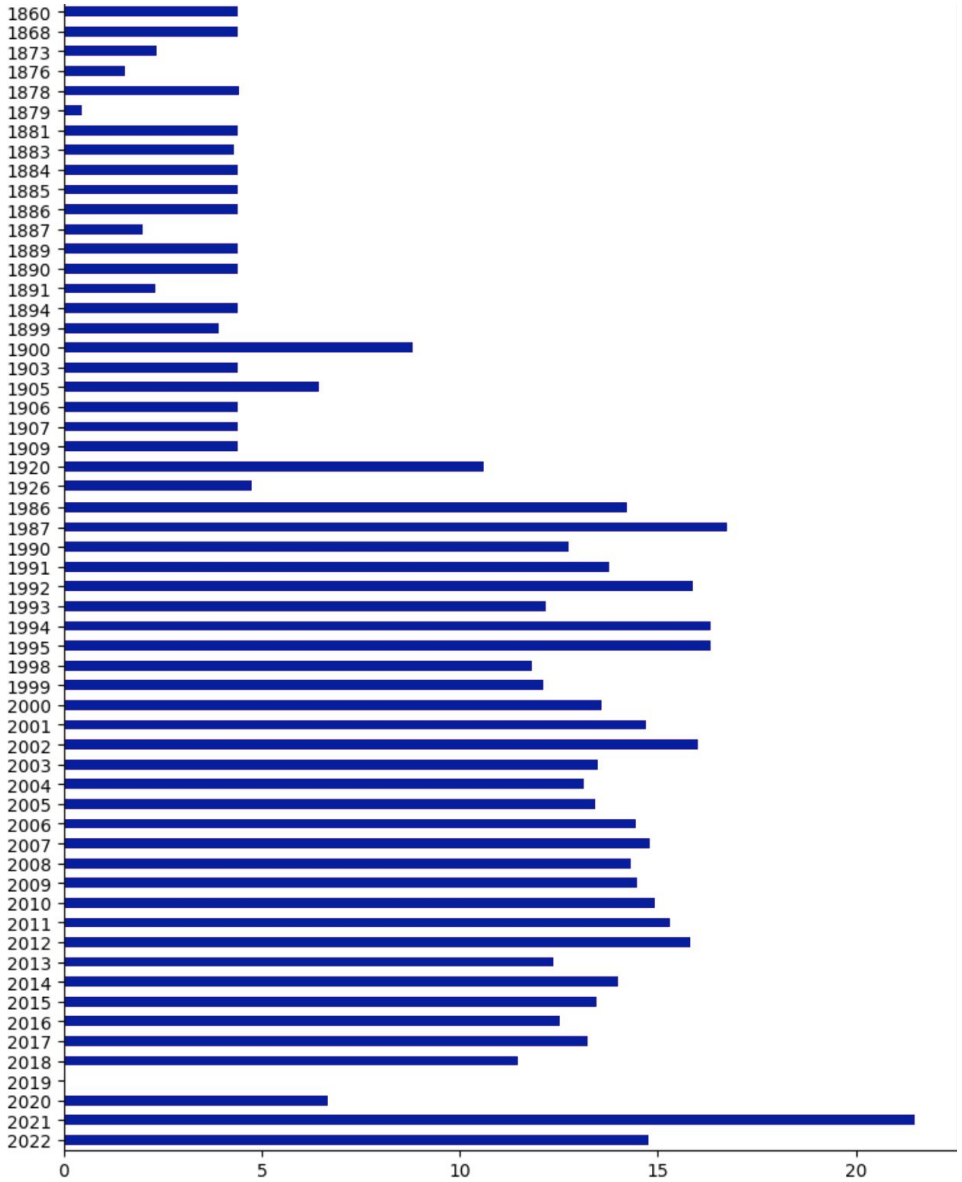
7 Analysis - Verb Pronominalization

Method: Take in (word, POS) pairs, look for:

Enclitic

Proclitic




1860
1868
1873
1876
1878
1879
1881
1883
1884
1885
1886
1887
1889
1890
1891
1894
1899
1900
1903
1905
1906
1907
1909
1920
1926
1986
1987
1990
1991
1992
1993
1994
1995
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
0
5
10
15
20

THANK YOU!

Any questions, comments or feedback will be greatly appreciated!

8 References

Montoya Abat, B., & Mas i Miralles, A. (2013). *La variació lingüística a la Governació d'Oriola*. **Treballs de sociolingüística catalana**, (23), 103–115. https://raco.cat/index.php/TreballsSocioling/article/view/281158

3Cat. (2022, 15 de juny). *El futur del català a "Sense ficció": el català, en perill d'extinció?*. https://www.3cat.cat/3cat/el-futur-del-catala/noticia/3160173/

Carme Junyent. (2023, 9). *El català és una llengua en procés d'extinció.* La Directa. https://directa.cat/el-catala-es-una-llengua-en-proces-dextincio/