Studying Variation and Change in Progress with the Penn Format

Eric Haeberli[1], Manuela Schönenberger[1], & Richard Zimmermann[2]
([1]University of Geneva, [2]University of Manchester)

UNIVERSITÉ DE GENÈVE
FACULTÉ DES LETTRES
Département de linguistique

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

MANCHESTER 1824
The University of Manchester

DiGS 25 – Universität Mannheim
25 June 2024

Background (i)

- The Penn corpora and their relatives have been highly successful in allowing us to gain important new insights into how a wide range of syntactic and morphosyntactic phenomena have evolved over time and more generally into the nature of linguistic variation and its interaction with change.
  - Lightfoot (2019a: 395): They have "revolutionized work in diachronic syntax".
  - Lightfoot (2019b: 331): They "have revolutionized our capacity to test hypotheses about, say, thirteenth-century I-languages".
- But the corpora inevitably suffer from limitations that are inherent to historical sources: Restricted to what has survived the accidents of history; evidence that would be crucial for a fuller understanding of certain phenomena may be missing; uncertainty to what extent the written sources reflect the spoken language of the time etc.

Background (ii)

- Work in the tradition of Labovian variationist sociolinguistics has shown how some of these problems can be addressed and how more detailed insights into the nature of language variation and change can be gained.
- Instead of sources from the past: Use of present-day data obtained through the selection of informants according to sociolinguistic criteria.
- Change over time is investigated by using the apparent-time method (and, potentially, subsequent real-time evidence).
- But few resources are available that allow variationist analysis of spoken vernaculars, and in particular their syntax.

Syntactic variation in a variety of Swiss German

- General aim of the project:
  - To gain a better understanding of inter-speaker and intra-speaker variation in syntax/morphosyntax on the basis of variation in one variety of Swiss German.
  - To gain a better understanding of the interaction between syntactic variation and change.
- Advantages of a Swiss German dialect for the study of variation:
  - A considerable number of aspects of the (morpho)syntax show variation.
  - Not a low-prestige variety: Used naturally in everyday life by members of all social classes (except in writing).
  - No normative pressure that could influence the use of variants in cases of variation.
  - Relatively resistant to influences of the standard.

Wilko - A parsed corpus of Swiss German

- A new relative of the Penn corpora: ***WilKo*** – *Geparstes Korpus von Spontansprachdaten des Schweizerdeutschen der Stadt Wil*.
  - Tagged and parsed in the Penn format
- Production data from 62 speakers speaking the same dialect
  - local dialect of Wil (SG), 24 000 inhabitants
  - all data obtained from informal interviews of 90+ MIN (120h of audio-recordings, ca. 1.4 million words)
- Choice of speakers
  - must be a native speaker of the local dialect spoken in Wil (growing up and attending school in Wil)
  - can be classified into one of the age groups, different social backgrounds
- Age groups

young: aged 20–30 (18 speakers: 10f/8m) (average age: 24)
middle-aged: aged 45–60 (20 speakers: 9f/11m) (average age: 54)
+ 2 female interviewers
elderly: aged 70+ (17 speakers: 8f/9m) (average age: 77)
no group: (5 speakers: 4f (average age: 64) and 1m (aged 40–45))

Two case studies in variation and change

(not quite syntax yet…)

- Indefinite neuter article:

  es (Chind) or *e (Chind)* 'a child'
  (German: no variation: *ein (Kind)*)

- 1sg present tense of *go* 'go':

  (i) gang or *(i) gò* 'I go'
  (German: no variation: *ich geh(e)*)

Aims

- Our initial aim was to identify linguistic or non-linguistic factors that may have an influence on the use of the two variants in each case.
- Here our main focus will be on two particular variables:
  - Age (potential apparent time effects)
  - Clause type

Clause type and language change (i)

- It has frequently been suggested in the literature that innovative properties first emerge in main clauses and that subordinate clauses tend to be more conservative (cf. Bybee 2002 for an overview).
- This contrast between clause types was initially identified in the context of word order change (cf. e.g. Hock 1986, Venneman 1975).
- But also in other contexts: E.g. subjunctive forms in certain languages have been argued to be the result of the emergence of a new morphological form in main clauses and the retention of an older form in subordinate clauses (Klein-Andreu 1991, Bybee et al. 1994)

Clause type and language change (ii)

- For word order changes or the emergence of the subjunctive, discourse-pragmatic factors have plausibly been invoked to account for the clause type contrast.
- Hypothesis: The relevant changes take place more readily in main clauses because of the more complex pragmatic relations and content of main clauses (cf. Bybee 2002, Matsuda 1998). Innovation arises from the reinterpretation of marked discourse variants as the neutral pattern.

Clause type and language change (iii)

- But: Such an account cannot be extended to all cases of clause type differences observed in the literature.
- Matsuda (1993, 1995): Two morphological changes in progress in Tokyo Japanese where the innovative option is more advanced in main clauses.
  - Replacement of a verbal suffix expressing potentiality
  - Omission of the accusative case morpheme
- For these contrasts, no plausible discourse-pragmatic explanation can be given (Matsuda 1998).
- Matsuda's observations seem to be somewhat unique in the literature on clause type differences in diachrony.
- Aim: To examine the two Swiss German variables presented above with respect to the potential interaction between morphological change and the syntactic factor of clause type.

Methods

- The data were retrieved from the unannotated version of the complete *Wilko* and coded manually.
- Statistical methodology:
  - The data distribution of both case studies was modeled with a number of logistic regression models predicting the proportion of one form from a number of factors.
  - Many models are possible. However, using trial and error, model evaluation metrics and hypothesis-guided common sense, one "best" model is proposed to capture the tendencies in the data.
  - The findings are evaluated with reference to this "best" model.

1st study: 1sg of *go* 'go' (*gang* vs. *gò*)

- There is variation between *gang* and *gò* for the first person singular form of the verb *go*.

	Singular (SG)		Plural (PL)	
1st	(i)	gang/gò	(me)	gönd
2nd	(du)	gòsch	(er)	gönd
3rd	(er/si/es)	gòòt	(si)	gönd

Table 1: Paradigm of *go* (present tense) in the Wil dialect of Swiss German.

- In the *Wilko*, there are 652 occurrences of the verb *go* in the first person singular: 464 *gang* (71.2%) vs. 188 *gò* (28.8%).

Independent variables

- The role of three independent variables is examined:
  - *Year of birth*: Apparent time.
  - *Clause type*: The finite verb does not occupy the same position in a matrix clause and in an embedded clause.

    (1a) Hüt gang/gòn i i d Schtadt. (V2: non-subject-initial)

    (1b) I gang/gò hüt i d Schtadt. (V2: subject-initial)

    (2) I ha der dòch gsait, dass i hüt i d Schtadt gang/gò. (VE: verb-final).
  - *Type of* go: Two types of *go* can be distinguished: main verb *go* and doubling verb *go*.

    (3a) Etz gang/gòn i mit de ÖV. 'Nowadays I go by public transport.'

    (3b) I gang/gò bald go raise. 'I'll soon go travelling.'

Results (i)

- Year of birth is significant; younger speakers use more *gò* (vs. *gang*).
  - The use of *gang* is in decline. At this rate of change, *gò* would reach a predicted probability of 99% of use with speakers born in the year 2107.
    - Year was scaled because it occurs on a large scale making it hard to estimate variances.
- Clause type is significant; main clauses have an overall higher probability of *gò* (vs. *gang*) than embedded clauses.
  - There is an important interaction between Clause type and Year of birth: As embedded clauses catch up with main clauses, the use of *gò* (vs. *gang*) increases at a faster rate of change.

Results (ii)

- Doubling is significant; the finite form of *go* in the doubling verb construction is overall more likely to be realized as *gò* (vs. *gang*).
- There are clear preferences among speakers to use one or the other form, necessitating the inclusion of random speaker intercepts.
  - Intraclass Correlation Coefficient (ICC): 94%, suggesting substantial variability between individuals.

Model illustration

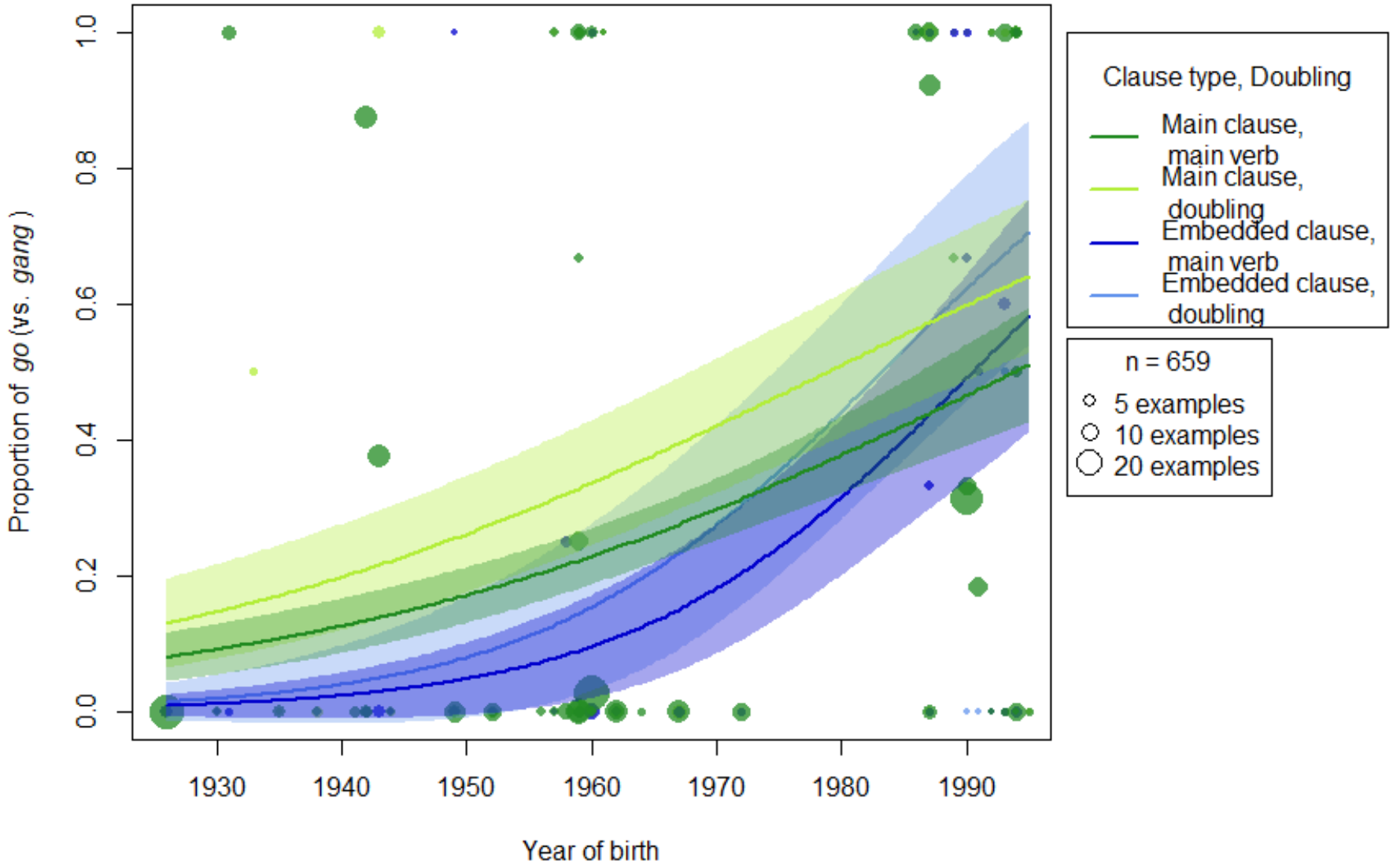

Figure 1: Logistic regression model predicting the frequency of *gò* (vs. *gang*) from year of birth, clause type and doubling.

Discussion (i)

- The *gang*/*gò* variable seems to represent a change in progress, with *gò* replacing *gang*.
  - The younger the speakers are, the more they use *gò.*
- The use of the two variants shows properties that have been related to change:
  - Main clauses are innovative, subordinate clauses are conservative: Higher frequency of *gò* in matrix clauses than in embedded clauses in the early stages of the change.
  - More frequently occurring items are more resistant to change than less frequently occurring items (cf. e.g. Bybee 2006): In doubling-verb contexts (n=121), the innovative form *gò* is more frequent than in main verb contexts (n=531).

Discussion (ii)

- At first sight, the developments with respect to clause type shown in Figure 1 seem to violate the Constant Rate Hypothesis. But:
  - With 652 observations, the sample size remains small.
  - We have not seen the complete change yet.
    - Zimmermann (2023: 332): "The CRH can be tested most rigorously with a syntactic change that is attested from beginning to end. Since any stage of a grammatical development can be influenced by unforeseen factors, omitting some intervals of the transitional period may result in incorrect conclusions about the potential constancy of a contextual effect".
  - The CRH as the null hypothesis.
    - Zimmermann (2022: 347): "it may be time to abandon the special status ascribed to the CRH as prima facie less plausible than its alternative and instead to treat it as a fallback position in the absence of evidence like any other null hypothesis. Once the CRH is viewed more widely as an ordinary null hypothesis, linguists should provide explanations when grammatical contexts do not develop at identical rates (language–internal interactions, cultural shifts etc.) rather than presuppose that rates of change may and will vary by default.

2nd Study: Indef. neut. article: *e* vs. *es*

- There is variation in the expression of the singular indefinite neuter article between the forms *e* and *es* in the nominative/accusative case.
- A linking *n* is often used between two words, one ending in a vowel and the other beginning with a vowel, to avoid a hiatus

(4) ˌ'**en** **A**maise vs. ??**e** **A**maise 'an ant$_{fem.}$'

	masc.	fem.	neuter
non-oblique NOM/ACC	en Maa 'a man'	e Frau 'a woman'	e/es Chind 'a child' en/es **A**uto 'a car'
prepositional P+ACC	in en Ruum 'into a room'	in e Buude 'into a stall'	in e/es Kino 'into a cinema' in en/es **A**uto 'into a car'
prepositional P+DAT	im ene Ruum 'in a room'	in ere Buude 'in a stall'	im ene Kino 'in a cinema'

Table 2: Partial paradigm of the indefinite article in the Wil dialect of Swiss German.

Indefinite neuter article: *e(n)* vs. *es*

- The use of *e* is nearly systematic in combination with *bitzeli* 'bit$_{Dim}$', *paar* 'a few', *Wiili* 'while$_{Dim}$', *Zitli* 'time$_{Dim}$': n=561; 541 *e(n)* (96.4%) – 20 *es* (3.6%).
  - These "formulaic" expressions generally occur only in the NOM/ACC, and they are not modified by adjectives.
- Excluding "formulaic" expressions from the variable context, we find 3,968 instances of the indefinite neuter article in NOM/ACC: 2,872 *e(n)* (72.4%) – 1,096 *es* (27.6%).

Independent variables

- The role of two independent variables is examined:
  - *Year of birth*: Apparent time
  - *Clause type*

Results (i)

- The data distribution represented by a smoothing trendline does not suggest a clear, important development in the use of *e* vs. *es* over time in different clause types (main, subordinate, fragments).

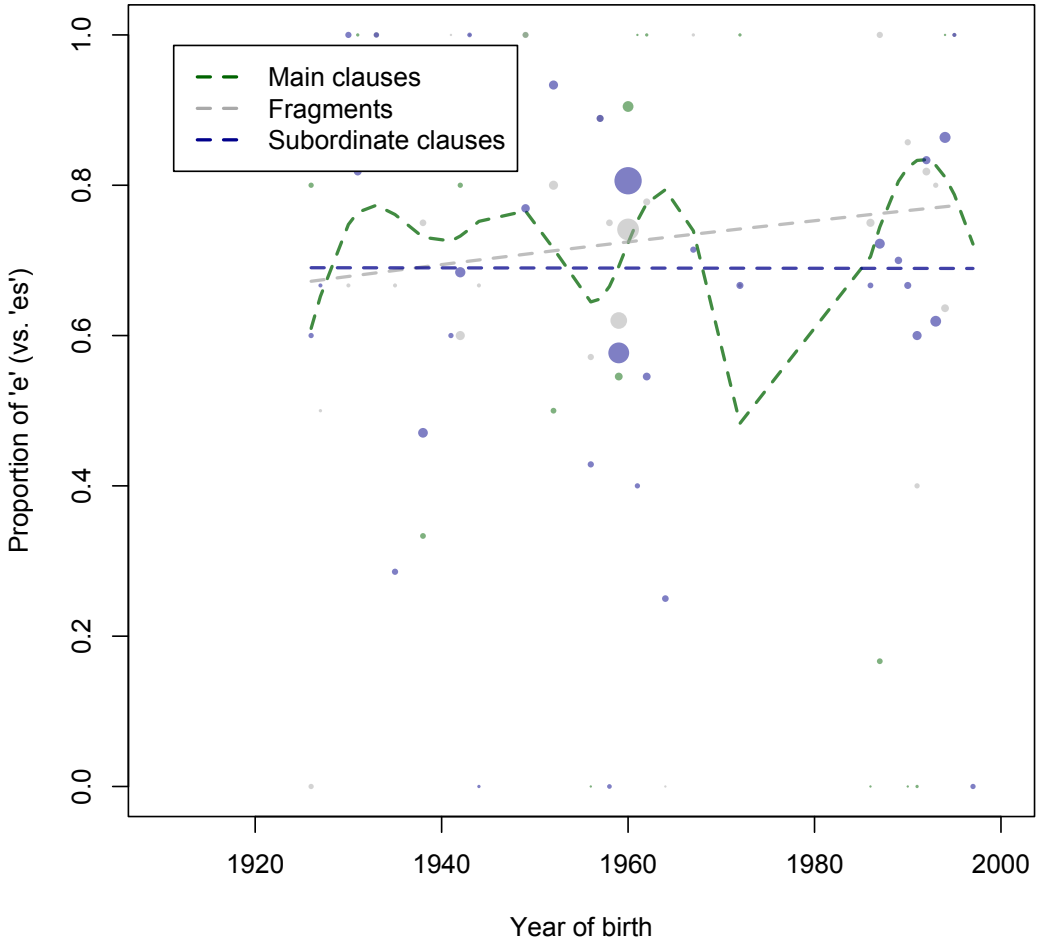

Figure 2: Distribution of *e*/*es* across year of birth and clause type with smoothing trends from a gam (generalised additive model) with a spline for year of birth.

Results (ii)

- In a mixed effects model that controls for clause type (subordinate vs. main/fragments) and takes into account clustering within the speaker, there is no significant time effect at all, and *e* vs. *es* remains stable both in main and subordinate clauses.

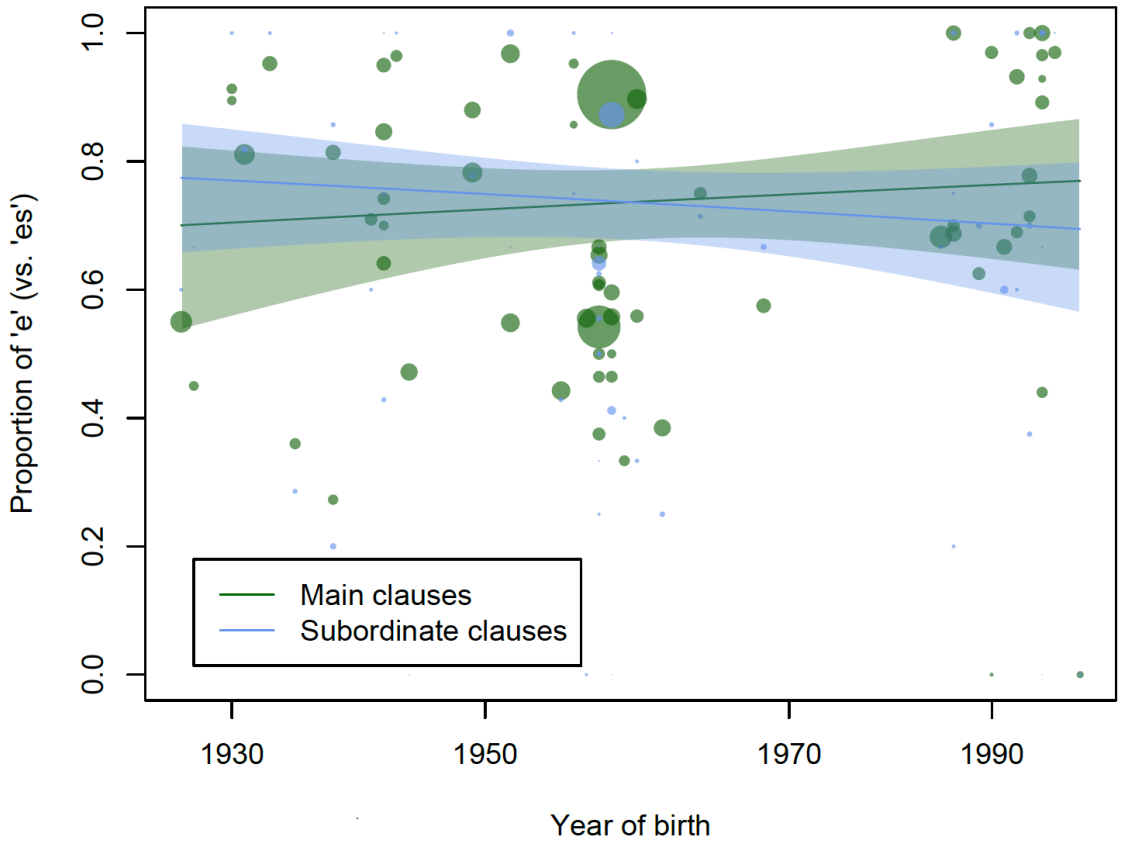

Figure 3: Logistic regression model predicting the frequency of *e(n)* (vs. *es*) from year of birth and clause type with random intercepts for speakers.

Discussion

- The variation with respect to the indefinite neuter article does not show any evidence for an apparent time effect and therefore seems to be relatively stable over time.
- With this case of variation no clause type effect can be identified.
- But: Other independent variables seem to influence the use of *e/es* as well (in particular the occurrence of an adjective between the determiner and the noun, and the following sound (sibilant vs. other)). The absence of an age and clause type effect will therefore have to be confirmed by a model taking these factors into account as well.

Conclusions (i)

- The nature of the two phenomena examined seems different:
  - 1 sg *gang/gò*:
    - A clear apparent-time effect suggesting a change in progress.
    - The innovative form is initially more advanced in main clauses than in subordinate clauses.
  - Indefinite neuter article *e(n)/es*:
    - No clear apparent-time effect, probably no change in progress.
    - The use of the two variants does not seem to be influenced by clause type.

Conclusions (ii)

- Our results are in line with Matsuda's (1993, 1995) finding that morphological change may show clause type effects. At the same time, they suggest that there is no such effect when the variation is relatively stable.
- Two possible ways to account for these findings, to be explored in further research:
  - Matsuda (1998): Processing factors may play a role in the way change spreads - related to psycholinguistic findings suggesting that subordinate clauses are more difficult to process than main clauses.
  - A range of interacting factors may lead to the contrast observed in our data, including possibly the type of elements involved in the variation (finite verb with *go*, DP-internal element with indefinite *e*/*es*).

Merci vilmòl !

References

Bybee, J. 2002. Main Clauses Are Innovative, Subordinate Clauses Are Conservative. In J. Bybee and M. Noonan (eds.), *Complex Sentences in Grammar and Discourse: Essays in Honor of Sandra A. Thompson*. Amsterdam: Benjamins. 1-17.

Bybee, J., R. Perkins, and W. Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World.* Chicago: University of Chicago Press.

Hock, H. H. 1986. *Principles of Historical Linguistics*. Berlin: Mouton de Gruyter.

Klein-Andreu, F. 1991. Losing Ground: A Discourse-Pragmatic solution to the history of *-ra* in Spanish. In S. Fleischman and L. R. Waugh (eds.), *Discourse-Pragmatics and the Verb: Evidence from Romance*. London: Routledge. 164-178.

Lightfoot, D. 2019b. Transparency. In A. Ledgeway and I. Roberts (eds.), *The Cambridge Handbook of Historical Syntax*. Cambridge: Cambridge University Press. 322-337.

Lightfoot, D. 2019b. Acquisition and Learnability. In A. Ledgeway and I. Roberts (eds.), *The Cambridge Handbook of Historical Syntax*. Cambridge: CUP. 322-337.

Matsuda, K. 1993. Dissecting Analogical Leveling Quantitatively: The Case of the Innovative Potential Suffix in Tokyo Japanese. *Language Variation and Change* 5: 101-133.

Matsuda, K. 1995. *Variable Zero-Marking of (o) in Tokyo Japanese.* Ph.D. Dissertation, University of Pennsylvania.

Matsuda, K. 1998. On the Conservatism of Embedded Clauses. *Theoretical and Applied Linguistics at Kobe Shoin* 1: 1-13.

Vennemann. T. 1975. An Explanation of Drift. In C. Li (ed.), *Word Order and Word Order Change*. Austin: University of Texas Press. 267-305.

Zimmermann, R. 2023. An Improved Test of the Constant Rate Hypothesis: Late Modern American English Possessive *have*. *Corpus Linguistics and Linguistic Theory* 19: 323-352.