The emergence of “weak” definite articles

Giuseppe Magistro[1,2]
Alexandra Simonenko[2]

[1] Urban/Eco, Università degli Studi di Napoli Federico II (Italy)
[2] Dept. of Linguistics, Ghent University, Belgium

25 September 2025
Sinn und Bedeutung 30
Goethe University Frankfurt

Table of contents

Two classes of definite articles

Some Germanic languages exhibit two morpho-phonological classes of definite articles: “weak” and “strong”. Since Ebert (1971) increasing attention has been devoted to this alternation.

(1) a. *Ik skal deel tu a / *di kuupmaan.*
I must down to DEF.W / *DEF.S grocer
‘I have to go down to the grocer.’

b. *Oki hee an hingst keeft. *A / Di hingst haaltet .*
Oki has a horse bought *DEF.W / DEF.S horse limps.
‘Oki has bought a horse. The horse limps.’

	m.Sg.	f.Sg.	n.Sg.	Pl.
A-article	*a*	*at*	*at*	*a*
D-article	*di*	*det*	*det*	*dön*

Table 1: A- vs. D-articles in Fering (Ebert, 1971)

Two classes of definite articles: the case of German

In Standard German, the alternation shows up in PPs: a full-fledged form vs. a prepositional enclitic, the former being restricted to contexts with an anaphoric antecedent (Schwarz, 2009, 2019).

(2) *In der Kabinettssitzung heute wird ein neuer Vorschlag*
in DEF cabinet.meeting today is a new proposal
vom *Kanzler / #Minister erwartet.*
by.DEF.W chancellor / minister expected
'In today's cabinet meeting, a new proposal by the chancellor/minister is expected.'

(3) a. *Hans hat gestern einen Minister interviewt.*
Hans has yesterday a minister interviewed
'Hans interviewed a minister yesterday.'

b. *In der Kabinettssitzung heute wird ein neuer Vorschlag* ***von***
in the cabinet.meeting today is a new proposal by
dem *Minister erwartet.*
DEF.S minister expected
'In today's cabinet meeting, a new proposal by the minister is expected.'

Research questions

Coniglio and Schlachter (2014): The use of reduced forms for weak articles was rare in Old High German and mostly constrained to NPs whose referents *"have to be unique"* (sic), e.g. *zur helle* 'to the Hell' or *ans cruce* 'to the Cross').

- RQ1: When did weak definite articles become more frequent?
- RQ2: Why do they emerge under such conditions?
- RQ3: Is there a rationale behind the mapping between morphophonological properties and the distributions?
- RQ4: Why do weak articles *do not* spread onto anaphoric contexts?

$\rightsquigarrow$ we will address these RQs with a combination of a) corpus data, b) computational pragmatics, and c) reinforcement learning agent-based simulations.

The emergence of weak articles

Determiner types in PPs in historical High German, based on IPCHG corpus (Sapp et al., 2024): 55 texts from 1250 to 1650.

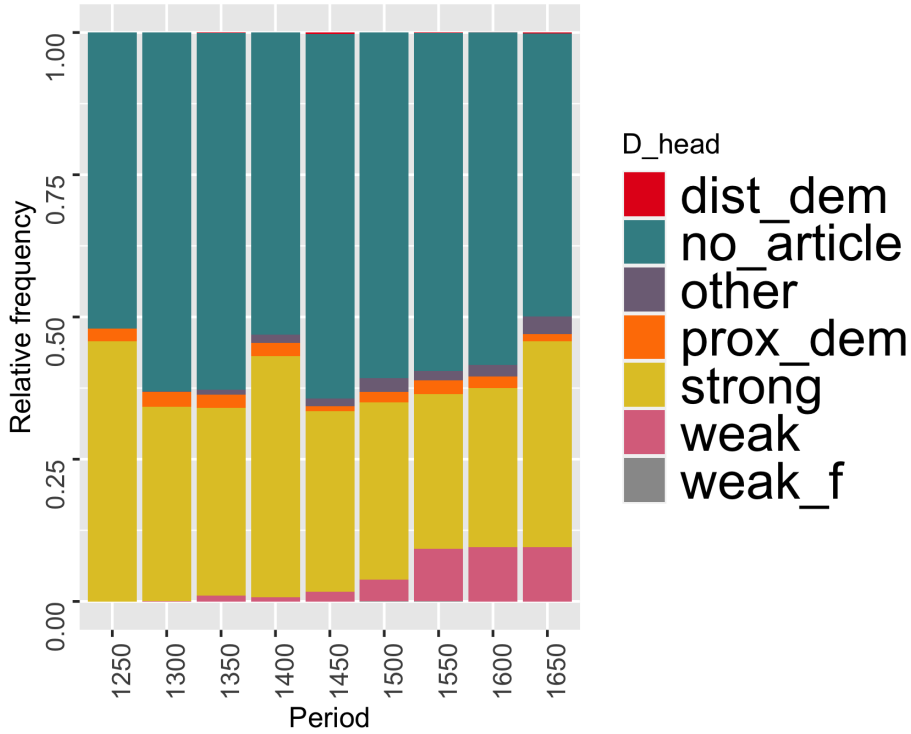


Figure 1: The emergence of weak articles in PPs in IPCHG

First occurrences of weak articles

... with NPs whose extension, in (almost) every situation in which there is one, **is a singleton** (cf. global situation uniqueness in Hawkins 1978).

(4) *Do der kuonik der eren vnser here ihenc xpc*
when the king DEF.GEN glory our lord Jesus Christ
zvr helle *qam [...]*
to.DEF.W hell came
"When the King of Glory, Our lord Jesus, went to hell [...] "
(Leipziger Rel-Sermon ThurSax. 77, 1300)

(5) *sie sach ein swester die waz urschaiden auz irm closter*
she saw a sister who was separated from her monastery
im paradis.
in.DEF.W paradise
She saw a sister who was away from the monastery, in paradise.
(Engelthaler Rel-Narrative 568, 1340)

A progressive diffusion of the weak definite articles

... to NPs for which the **probability of having a singleton extension in a randomly chosen situation is (intuitively) lower**.

(6) *Da sagt yn myn herre Gawan wie er* ***zum*** ***karch***
Then said him my lord Gawan how he to.DEF.W church
komen was [...]
come was
"Then my lord Gawan told him how he had come to the church [...]" *(Karrenritter Lit-prose 207, 1430)*

(7) *Er qwam wiedder* ***zum*** ***bette*** *vnd die Jungfrau fraget yn*
He came back to.DEF.W bed and the maiden asked him
was er thun wolt
what he do wanted
"He came back to bed and the maiden asked him what he wanted to do" *(Karrenritter Lit-prose 239, 1430)*

Core assumptions: Part I

- Articles **guide reference resolution** by indicating in what kind of situation the relevant extension is to be found:
  - **morphophonologically fully-fledged forms** signal that the hearer should search for **a situation with a unique individual** satisfying the NP-property;
  - **reduced forms are ambiguous** in being compatible with situations having a unique as well as multiple individuals satisfying the NP-property.
- Speakers tend to **avoid ambiguity**...
- But they also tend to **produce hypo-articulated forms** (Cangemi and Niebuhr, 2018; Clopper and Turnbull, 2018) (e.g. *von əm*).

The article choice is a matter of **tension between economy and clarity**.

Core assumptions: Part II

- Each NP type is associated – as a matter of mutually assumed background world/lexical knowledge – with **a specific probability of having a singleton extension in a randomly chosen situation**.
- A speaker is more likely to try and **"get away" with an ambiguous article** form provided the *apriori* probability of the relevant NP having a singleton extension is **high enough** (to be defined).
- But in anaphoric contexts, the corresponding likelihood goes down:
    - **An indefinite antecedent makes it less likely** that a given NP has a singleton extension in a randomly chosen situation.

Computational modeling: the goals

- We expect NPs with a relatively high probability of having a singleton extension in any situation to be relatively more prone to combining with reduced articles.
- To test this, we simulate the (evolving) use and the resulting distribution of the two forms using computational simulations, involving a series of referential games between two rational agents.

Computational modeling: the game set-up

Each agent is endowed with a series of attributes:

- a randomly assigned **age** until 80;
- a randomly assigned **lifespan** (number of games);
- a rationality parameter $\alpha \in [1, 2]$ which controls the agent's **clarity**;
- **economicity** = cost penalizing redundant forms;
- **conservativity** = cost penalizing novel forms.
- a **learning rate** $\eta \in [1, 2]$ that updates the costs according to the success/failure of a referential game (for players younger than 12, $\eta = 2$ to account for changes in L1 acquisition).

Computational modeling: the game set-up

- **Two information states that can be conveyed**: $s_1 = \{e_1\}$ and $s_2 = \{e_1 \ ... \ e_n\}$
  - $s_1 = \{e_1\}$: the intended **situation contains a single individual** with the relevant NP-property;
  - $s_2 = \{e_1 \ ... \ e_n\}$: the intended **situation contains multiple individuals** with the relevant NP-property.
- δ = **The probability of a given NP to have a singleton extension**, or *prior* probability of s_1, with $\delta \in [0, 1]$ (e.g. $\delta_{paradise} = 0.9$, $\delta_{church} = 0.5$)
- β = **The prior probability of a given NP to have an extension larger than a singleton**, or *prior* probability of s_2, ($\beta = 1 - \delta$).
- **Two article forms**: u_1 (the fully fledged form) and u_2 (the reduced form)
  - u_1 can only express s_1;
  - u_2 can express either s_1 or s_2.

Computational modeling: the game set-up

- **Two contexts**: anaphoric (an identical NP_2 in the previous clause) or non-anaphoric contexts (the rest).
- **In anaphoric contexts, δ is overwritten by $\delta' = 1 - \beta * \gamma$:**
  - γ: the probability that NP_1 is co-referential with NP_2 (set to 0.9);
  - β: the probability that NP_2 has an extension larger than a singleton in a random situation ($\beta_{NP_1} = \beta_{NP_2}$);
  - $1 - \beta * \gamma$: the likelihood of NP_1 having a singleton extension given the probability of a preceding co-referential indefinite NP_2.
- for "paradise", $\delta = 0.9$, $\delta' = 0.81$ (an indefinite antecedent is very unlikely, so δ and δ' are very close);
- for "church", $\delta = 0.5$, $\delta' = 0.05$ (an indefinite antecedent is much more likely, so δ' is much lower than δ).

Computational modeling: the game dynamics

- Either an anaphoric or non-anaphoric context is assigned to the game.
- Each player is randomly assigned a role (either the speaker or the listener).
- The speaker is assigned a random state (either s_1 or s_2) and must guide the listener by using either form (u_1, u_2).
- The speaker's choice is modeled according to the Rational Speech Act framework (Frank and Goodman, 2012)
  - Each participant is a rational Bayesian agent who will simulate the behaviour of the other interlocutor and act upon probabilistic reasoning.
  - Using the Python library Pyro (Bingham et al., 2019), each participant estimates the posterior probability distribution of the utterance/state (depending on their role of speaker/listener).

Computational modeling: the game dynamics

- Using δ, α and the different costs for the two forms, the pragmatic speaker simulates the behaviour of a literal listener (who leverages a literal mapping between article forms and information states) and decides on the appropriate form (Goodman and Frank, 2016).

$$U(u; s) = \alpha \cdot \log P_{L_0}(s \mid u) - C(u).$$

$$P(s \mid u) \propto \delta$$

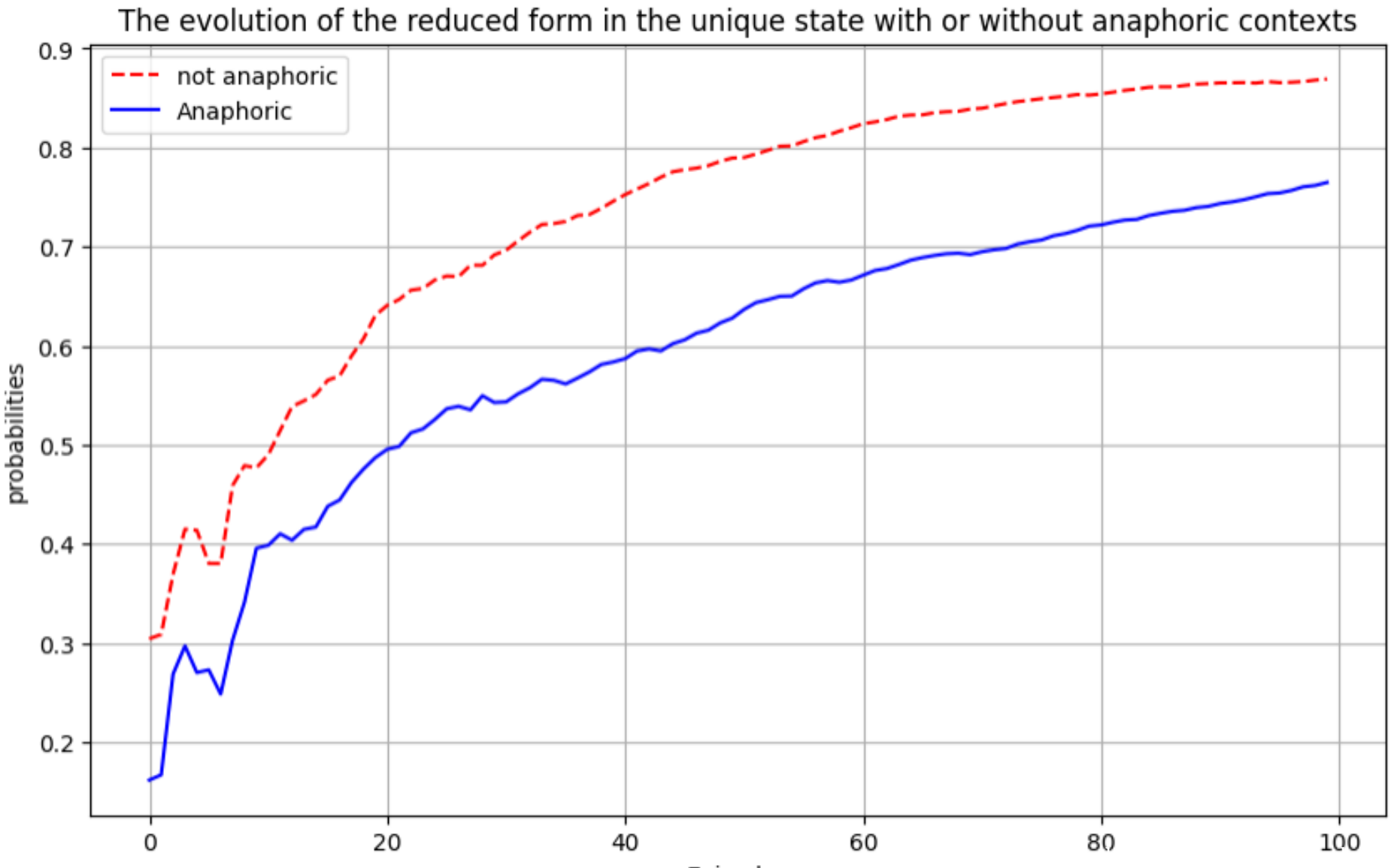
- Once the form u has been chosen by the speaker, the listener receives u as input and simulates the reasoning of a pragmatic speaker, to estimate the posterior probability of the state s and decide which state to infer.
- If the state chosen by the listener corresponds to the one initially assigned to the speaker, **the speaker's costs for the selected form are reduced** by $c' = c - \eta$ and **the alternative form is penalized with a cost increase** ($c' = c + \eta$).

Computational modeling: the population

- After each game, each player gets older by 1 unit.
- When a participant "dies", they are replaced by a new young participant (with $\eta = 2$, who inherits the costs from their parents).
- For each run, we had 40 interacting agents in couple.
- 200 total episodes (100 in anaphoric, 100 in non-anaphoric contexts)
- We ran the 200 episodes for two δ values: $\delta = 0.9$ (for nouns such as "paradise"), $\delta = 0.5$ (e.g. "church")
- Total games $= 40 * 100 * 2 = 8000$
- For each episode the mean probability distribution of each form given each state was computed and plotted.

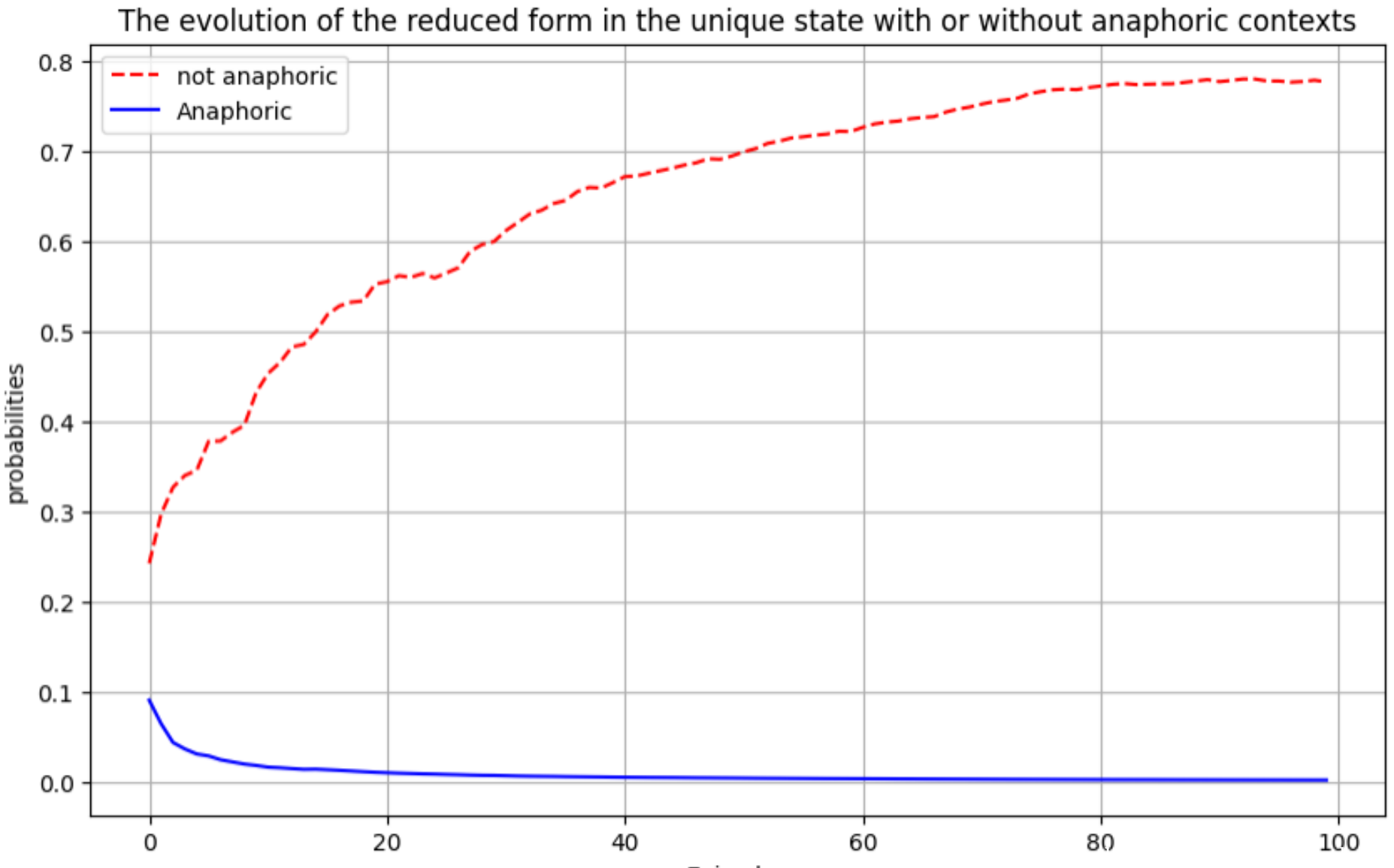
Results: nouns with high δ

$\delta = 0.9$ (e.g. “Vatican”, “Paradise”, “Moon”)



Results: nouns with low δ

$\delta = 0.5$ (e.g. “church”, “dog”, “chair”)



Discussion and conclusions

- In non-anaphoric contexts, the weak form spreads rapidly given a "unique" state (i.e. the target situation contains a single individual with the relevant NP property), especially for NPs with a high probability of having a singleton extension.
- Anaphoric contexts display a contrast which depends on δ: while NPs with high δ will also tend to being introduced by weak articles, NPs with lower δ will never occur with a weak form.
- We were able to reconstruct the spread of weak forms to express the "unique" state and the eventual constraining of the strong forms to the anaphoric contexts.
- Assuming different probabilities of having a singleton extension together with principles such as clarity and economy can help reconstruct the change that we see in corpora.

Thanks!

This project is co-funded by the European Union (ERC, CAUSALITY, 101042427). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References I

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6.

Cangemi, F. and Niebuhr, O. (2018). Rethinking reduction and canonical forms. In Niebuhr, O., Barbara Schuppler, F. C., Clayards, M., and Zellers, M., editors, *Rethinking Reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*, pages 291–316. De Gruyter, Berlin.

Clopper, C. G. and Turnbull, R. (2018). Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. In Cangemi, F., Clayards, M., Niebuhr, O., Schuppler, B., and Zellers, M., editors, *Rethinking Reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*, pages 39–86. De Gruyter, Berlin.

Coniglio, M. and Schlachter, E. (2014). Referential properties of the full and reduced forms of the definite article in german: A diachronic survey. In *Information Structure and Syntactic Change in Germanic and Romance Languages*, pages 141–172. John Benjamins Publishing Company.

References II

Ebert, K. H. (1971). *Referenz, Sprechsituation und die bestimmten Artikel in einem nordfriesischen Dialekt*, volume 4. Nordfriisk Instituut, Bredstedt.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Hawkins, J. A. (1978). *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. Croom Helm, London.

Sapp, C. D., Evans, E., Sprouse, R., and Dakota, D. (2024). Introducing a parsed corpus of historical High German. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9224–9233.

Schwarz, F. (2009). *Two Types of Definites in Natural Language*. Ph.d. dissertation, University of Massachusetts Amherst, Amherst, MA.

Schwarz, F. (2019). Weak vs. strong definite articles: meaning and form across languages. In *Definiteness across languages*. Language Science Press.