3SG is the most conservative subject marker across languages: An exploratory study of rate of change

Peter Dekker*[1], Sonja Gipper[2], and Bart de Boer[1]

*Corresponding Author: peter.dekker@ai.vub.ac.be
[1]AI Lab, Vrije Universiteit Brussel, Brussels, Belgium
[2]Institut für Linguistik, Philosophische Fakultät, Universität zu Köln, Cologne, Germany

In this paper, we investigate the rate of change of different person-number subject markers. We perform a cross-linguistic study on the dissimilarity between proto and modern forms, showing that 3SG is the most conservative subject marker across languages. We discuss the mechanisms that could explain this diachronic pattern, such as frequency of use, markedness, and attractor lengths. Our exploratory analysis highlights how existing linguistic datasets can be used to study new research questions.

1. Introduction

Many languages mark the person and number of the subject by means of a bound morpheme on the verb (Siewierska, 2013; e.g., *walk-s*, with *-s* marking the third person singular). These verbal person-number subject markers are known to change over time, with certain diachronic changes in paradigms of subject markers being more probable than others (Cysouw, 2001); for instance, the form for 3SG is more likely to extend to other persons than vice versa (Baerman, 2005). But it is less well studied whether there is a difference in the *rate of change* across the different person-number combinations. Are certain subject markers more prone to change than others? This is the question we set out to investigate in this paper.

We perform an exploratory quantitative study of rates of change for six different person-number combinations – first, second, and third person, each in singular and plural – in a sample of 310 languages (Seržant & Moroz, 2022, data publication: Seržant, 2021). We find that 3SG is the most conservative subject marker across languages. We then discuss these findings in light of possible factors that may be responsible for this pattern. We suggest that, in line with previous work (Pagel, Atkinson, & Meade, 2007; Hoekstra & Versloot, 2019), our data hint at an important role for frequency in the rate of change of subject markers, as it could plausibly be the driving factor in the pattern we observe, while also relating to other possible explanations such as markedness and attractor lengths.

2. Method

Our sample of 310 modern languages associated with 15 proto-languages constitutes a subset of data[1] from Seržant (2021) who created a sample of subject markers in 383 languages from 53 families, as well as the reconstructed forms in their respective proto-languages, for six grammatical persons: 1SG, 2SG, 3SG, 1PL, 2PL, 3PL.[2] We calculate the Levenshtein distance (Heeringa, 2004) between proto and modern forms.[3] We use this degree of dissimilarity between proto and modern forms as a proxy for *rate of change*, i.e., amount of change over time period. Given the uncertainties regarding the estimation of the age of language families (Maurits, de Heer, Honkola, Dunn, & Vesakoski, 2020), our approach is agnostic with respect to the potentially different ages of the language families and proto-languages.[4]

The results of our distance calculation depend on the reconstructions of proto-forms, about which there is not always a consensus or which might represent an abstraction. Therefore, when comparing proto to modern forms, we do not assume that these comparisons necessarily represent concrete changes with historical reality. Rather, we aim to search for a general signal of cross-linguistic differences between subject markers. Moreover, reconstructed proto-forms generally give an underestimation of change, as traits of the proto-language not preserved in the daughter languages are not included in the reconstructed form (Campbell, 2013, p. 144). Despite these remaining uncertainties, we think this comparison between proto-forms and modern forms can serve as a fruitful first exploration of our research question, sketching an approach to explore an existing cross-linguistic dataset to find evidence for a novel linguistic question (cf. Ladd, Roberts, & Dediu, 2015).

We analyse the data using a mixed linear model (details in SI). The Levenshtein distance constitutes the response variable and person and number serve as

[1] All code of this paper can be found in `https://zenodo.org/doi/10.5281/zenodo.10722183` and the GitHub repository `https://github.com/peterdekker/changesubjectmarkers`. The Supplementary Information of this paper contains additional information on the technical details of the applied method.

[2] The dataset does not report forms that show a contrast in terms of clusivity, nor dual or paucal subject markers.

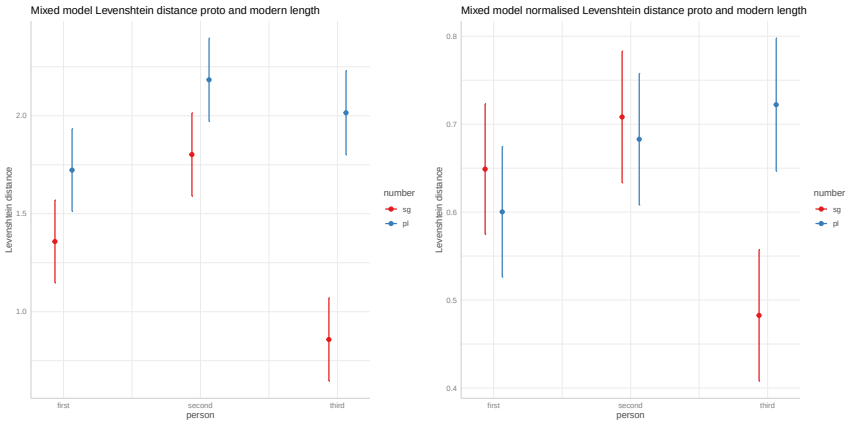
[3] For this exploratory analysis, we calculate the distance between orthographic forms as reported in the dataset. A more fine-grained analysis could be conducted in the future by using phonetic forms or even taking into account phonetic features (cf. List, 2012; Mortensen et al., 2016).

[4] Assuming that any specific age would apply in the same way to all person markers of a given language, we propose that family age can be neglected in our analysis. Also, Rama and Wichmann (2020, Table 6) show that family ages are in the same order of magnitude, for a sample overlapping ours. Moreover, in general the age of proto-languages is bounded by the time depth of reconstruction of the comparative method: maximum 6,000–10,000 years (Campbell, 2013, p. 341). For a more precise treatment of proto-language age, one could include a phylogenetic model in the analysis (e.g. Hahn & Xu, 2022).

predictors, with an interaction between person and number. We use clade as a random effect, because data points from languages in the same clade in a family should be treated as not fully independent, even more so because the Levenshtein distances are calculated with respect to the same proto-language. This random effect also partially addresses the potentially different ages of proto-languages. We report normalised and unnormalised Levenshtein distance. Unnormalised Levenshtein distance corresponds to a theory of a fixed rate of change *per form*: every timestep, there is a certain probability that 1 segment in the form will change. Whatever the length of the form, a change of 1 segment gives a Levenshtein distance of 1. On the other hand, normalised Levenshtein distance (distance divided by the length of the longest form), is based on a theory of a fixed rate of change *per phoneme* in a language. This assumes regular sound change, where a certain segment is substituted by another segment in all the forms in the language. For example, if in a language, the words ab and $abab$ have changed to ac and $acac$, due to the regular sound change $b \rightarrow c$, both receive a normalised distance 0.5, assigning the same score to forms affected by the same process of change. In this way, normalisation accounts for the fact that long forms have a higher chance of containing phonemes subject to regular sound change. Normalised Levenshtein distance is commonly used in phylogenetic reconstruction of language families (Serva & Petroni, 2008), which depends to a large extent on regular sound changes. For our purposes, to identify the rate of change per person marker, agnostic of the processes of change that are involved, we believe unnormalised Levenshtein distance is most suitable. However, we also report normalised Levenshtein distance for comparison.

3. Results

The predictions of the mixed linear model are shown in Figure 1. In the unnormalised model (Figure 1a), 3SG is the most conservative, while 2PL and 3PL are most innovative. Overall, singular forms are, on average, more conservative than their plural counterparts. The normalised (Figure 1b) model also shows 3SG as most conservative, while the difference between singular and plural can no longer be observed for first and second person. In sum, the most robust finding across both models is that 3SG is the most conservative among the six subject markers.



(a) Unnormalised Levenshtein distance (b) Normalised Levenshtein distance

Figure 1.: Predictions and 95% confidence intervals of a mixed linear model, with Levenshtein distance predicted from person and number (interaction), with clade as random effect. Higher values signify higher rates of change.

4. Discussion

We now turn to some possible explanations for our finding that in our analysis 3SG is the most conservative subject marker across languages. One factor that could arguably lead to this pattern is frequency of use, which has been shown to influence language change in at least two ways (Bybee & Thompson, 1997; Diessel, 2007, pp. 117–123; Hoekstra & Versloot, 2019): a conserving effect on morphology and a reducing effect on phonetics (Hinskens, 2011, p. 442). Both types of frequency effects are relevant for 3SG subject markers, as these tend to be both more conservative (our study) and shorter (Seržant & Moroz, 2022) than other subject markers.

Let us first turn to the conserving effect of frequency, based on the observation that high frequency of use reinforces the representation of a form, thereby preventing high-frequency irregular forms from becoming regularised (Diessel, 2007). Our finding that 3SG is the most conservative subject marker is consistent with this conserving effect of frequency, as there is evidence that in spoken language 3SG is the most frequent type of subject (e.g., Bybee, 1985, p. 71 on Spanish; Scheibman, 2001, p. 68, on American English; Seržant & Moroz, 2022, pp. 5–7, on Russian).

Regarding the second, reducing effect of frequency, it is also consistent with the rates of changes for the different persons presented in section 3: specifically the unnormalised model shows some parallels to the attractor lengths for the different persons reported in Seržant and Moroz (2022, Figure 2), which were cal-

culated on the basis of the same dataset.[5] Seržant and Moroz (2022) attribute the different attractor lengths to the reducing effect of frequency, with 3SG having the shortest attractor length.

The most extreme case of this is zero marking, which is cross-linguistically more common for 3SG than for other persons (Cysouw, 2001, pp. 53–58; Bickel, Witzlack-Makarevich, Zakharko, & Iemmolo, 2015, pp. 47–48). Moreover, proto-forms reconstructed as zero seem to be relatively conservative in our dataset.[6] Again, this is consistent with the finding that 3SG zero is more common in some families than others, i.e. that it is to some extent a genealogical phenomenon (see summary in Cysouw, 2001, pp. 53–58). A possible explanation for this conservative behaviour of 3SG zero in particular, at least in some cases, could be that some linguistic systems depend on 3SG to be zero-marked, such as in omnipredicative languages where all open lexical classes are basically predicates (Launey, 2004) – see Cristofaro (2021) for further possible factors that may lead to the non-development of a marker for 3SG. So possibly, the conservative nature of 3SG zero forms in combination with the generally low potential for change due to its short attractor length could explain our results instead of or in addition to frequency, although these factors relate to frequency.

Another factor that may have an influence on the rate of change in person markers is markedness. In a feature-based description of subject markers, it is generally assumed that the first and second person are more marked than the third person, as the latter does not exhibit the features of being a speech act participant and of being the author of the utterance (Buchler & Freeze, 1966, p. 81; Buchler, 1967, p. 42; Nevins, 2007). Furthermore, plural is more marked than singular (Cysouw, 2007, p. 6), which results in the lowest markedness for 3SG. In general, frequency and markedness go hand in hand, with marked forms also being less frequent (Bybee, 2010). Baerman (2005) suggests that markedness may explain the cross-linguistic tendency for it to be more likely that other persons (notably 1/2SG) take over the form of 3SG than vice versa. As the least marked and hence 'default' form, 3SG is more likely to extend to other persons. This is consistent with our finding that 3SG is the most conservative marker, as in this scenario, 3SG remains unchanged.

There are further aspects that will be necessary to integrate in a full investigation of rate of change in subject markers. For instance, it is clear that social dynamics impact on the rate of change of linguistic items, such as community size (Nettle, 1999). In addition, Cristofaro (2021) emphasises that in diachronic typol-

[5] However, our results for rate of change are not just an artefact of the lengths of the markers, as the normalised model (Figure 1b), where length of the person markers has largely been removed as a factor, still partially follows the patterns of the attractor lengths from Seržant and Moroz (2022), at least for the singular forms.

[6] For 3SG, in 82 out of 132 cases where the proto-form is zero, we observe a modern form that is also zero (62%).

ogy, it is important to take the different diachronic paths into account that can lead to a typological pattern. For our research question, this means that we should not only look at the overall rate of change of person markers, but also at the different diachronic paths that lead to more conservative 3SG. Moreover, it is necessary to tease apart frequency effects on rate of change from those on typologically preferred patterns. Cathcart, Herce, and Bickel (2022) present a study that suggests that frequency, rather than impacting on the *rate of change*, has an influence on *long-term preferences*, where more frequent lemmata in Romance languages are more likely to exhibit a stem alternation than less frequent ones. However, no influence of lemma frequency on rate of change was observed.

Finally, the role of processing in the change of subject markers and in language change in general will be a promising avenue for future investigation (see Bambini et al., 2021). There is some pioneering research on the effect of markedness in person agreement on online processing. In an ERP experiment, Alemán Bañón and Rothman (2019) find that in agreement mismatches in Spanish, there is a stronger P600 effect[7] when a 1SG subject is used with a mismatching 3SG verb, than in cases where a 3SG subject is combined with a mismatching 1SG verb. Such findings are highly relevant for investigating cases where the form of one person marker extends to other persons, but also for tendencies regarding rate of change across different person markers in general. Integrating the different strands of evidence will be an intriguing topic for future research.

5. Conclusion

In this paper, we showed an exploratory approach of using an existing linguistic dataset for a new research question. We found that 3SG is the most conservative subject marker and argued that frequency of use and, relatedly, markedness seem to be important factors influencing the rate of change of person markers. We would furthermore like to highlight the correlation with proposed cross-linguistic attractor lengths of different persons and the presence of zero markers. Further research will be necessary to tease these factors apart.

Acknowledgements

We are grateful to the reviewers of this article and an earlier submission for useful feedback. An earlier version of this work was presented to historical linguists at the University of Cologne; thanks to the audience for their comments. We would like to thank Yannick Jadoul for help with integrating R code into Python. All remaining errors are ours. PD was supported by a PhD Fellowship fundamental

[7] In recent research, it is argued that the P600 represents the degree of difficulty of integrating the target word in the unfolding utterance meaning (Aurnhammer, Delogu, Schulz, Brouwer, & Crocker, 2021; Aurnhammer, Delogu, Brouwer, & Crocker, 2023). The stronger the violation of expectations, the higher the P600 effect.

research (11A2821N) of the Research Foundation — Flanders (FWO). SG was funded by the University of Cologne Excellent Research Support Programme, funding line FORUM, project *Conversational Priming in Language Change*, as well as funding line Cluster Development Program, project *Language Challenges*.

References

Alemán Bañón, J., & Rothman, J. (2019). Being a participant matters: Event-related potentials show that markedness modulates person agreement in Spanish. *Frontiers in Psychology*, *10*, 746.

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, *60*(9), e14302.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*(9), e0257430.

Baerman, M. (2005). Typology and the formal modelling of syncretism. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 2004* (pp. 41–72). Dordrecht: Springer Netherlands.

Bambini, V., Canal, P., Breimaier, F., Meo, D., Pescarini, D., & Loporcaro, M. (2021). Capturing language change through EEG: Weaker P600 for a fading gender value in a southern Italo-Romance dialect. *Journal of Neurolinguistics*, *59*, 101004.

Bickel, B., Witzlack-Makarevich, A., Zakharko, T., & Iemmolo, G. (2015). Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories. In J. Fleischer, E. Rieken, & P. Widmer (Eds.), *Agreement from a diachronic perspective* (pp. 29–51). Berlin/Boston: De Gruyter.

Buchler, I. R. (1967). The analysis of pronominal systems: Nahuatl and Spanish. *Anthropological Linguistics*, *9*(5), 37–43.

Buchler, I. R., & Freeze, R. (1966). The distinctive features of pronominal systems. *Anthropological Linguistics*, *8*(8), 78–105.

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. In M. L. Juge & J. L. Moxley (Eds.), *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure* (Vol. 23, pp. 378–388). Berkeley.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* Amsterdam/Philadelphia: John Benjamins.

Bybee, J. L. (2010). Markedness: Iconicity, economy, and frequency. In *The Oxford Handbook of Linguistic Typology* (pp. 131–147). Oxford University Press.

Campbell, L. (2013). *Historical linguistics.* Edinburgh: Edinburgh University Press.

Cathcart, C., Herce, B., & Bickel, B. (2022). Decoupling speed of change and

long-term preference in language evolution: Insights from Romance verb stem alternations. In *Proceedings of the Joint Conference on Language Evolution (JCoLE)* (pp. 101–108). Kanazawa, Japan: JCoLE.

Cristofaro, S. (2021). Typological explanations in synchrony and diachrony: On the origins of third person zeroes in bound person paradigms. *Folia Linguistica Historica*, *42*(1), 25–48.

Cysouw, M. (2007). Building semantic maps: The case of person marking. In M. Miestamo & B. Wälchli (Eds.), *New challenges in typology: Broadening the horizons and redefining the foundations* (pp. 225–247). Berlin: Mouton de Gruyter.

Cysouw, M. A. (2001). *The paradigmatic structure of person marking.* Unpublished doctoral dissertation, Katholieke Univ. Nijmegen, Nijmegen.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*(2), 108–127.

Hahn, M., & Xu, Y. (2022). Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences*, *119*(24), e2122604119.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance.* Unpublished doctoral dissertation, Rijksuniv. Groningen.

Hinskens, F. (2011). Lexicon, phonology and phonetics. Or: Rule-based and usage-based approaches to phonological variation. In P. Siemund (Ed.), *Linguistic universals and language variation* (pp. 425–466). Berlin: De Gruyter Mouton.

Hoekstra, E., & Versloot, A. P. (2019). Factors promoting the retention of irregularity: On the interplay of salience, absolute frequency and proportional frequency in West Frisian plural morphology. *Morphology*, *29*(1), 31–50.

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annual Review of Linguistics*, *1*(1), 221–241.

Launey, M. (2004). The features of omnipredicativity in Classical Nahuatl. *STUF - Language Typology and Universals*, *57*, 49–69.

List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In M. Butt, S. Carpendale, G. Penn, J. Prokić, & M. Cysouw (Eds.), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 117–125). Avignon, France: Association for Computational Linguistics.

Maurits, L., de Heer, M., Honkola, T., Dunn, M., & Vesakoski, O. (2020). Best practices in justifying calibrations for dating language families. *Journal of Language Evolution*, *5*(1), 17–38.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. S. (2016). PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers* (pp. 3475–3484).

Osaka, Japan: The COLING 2016 Organizing Committee.

Nettle, D. (1999). Is the rate of linguistic change constant? *Lingua*, *108*(2-3), 119–136.

Nevins, A. (2007). The representation of third person and its consequences for person-case effects. *Natural Language & Linguistic Theory*, *25*(2), 273–313.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 717–720.

Rama, T., & Wichmann, S. (2020). A test of Generalized Bayesian dating: A new linguistic dating method. *PLOS ONE*, *15*(8), e0236522.

Scheibman, J. (2001). Local patterns of subjectivity in person and verb type in American English conversation. In J. L. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 61–89). Amsterdam/Philadeophia: John Benjamins.

Serva, M., & Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, *81*(6), 68005.

Seržant, I. (2021). *Dataset for the paper "Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved"*. https://zenodo.org/records/7641119: Zenodo.

Seržant, I. A., & Moroz, G. (2022). Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved. *Humanities and Social Sciences Communications*, *9*(1), 58.

Siewierska, A. (2013). Verbal person marking. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.