Number agreement, dependency length, and word order in Finnish traditional dialects

Sinnemäki, Kaius

The Association for Computational Linguistics
2021

Sinnemäki , K & Takaki , A 2021 , Number agreement, dependency length, and word order in Finnish traditional dialects . in Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021) . The Association for Computational Linguistics , Stroudsburg , pp. 115-123 , Workshop on Universal Dependencies , Sofia , 21/03/2022 .

http://hdl.handle.net/10138/338580

Depling 2021

Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)

Proceedings

To be held as part of SyntaxFest 2021
21–25 March, 2022
Sofia, Bulgaria

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-14-8

Table of Contents

Number agreement, dependency length, and word order in Finnish traditional dialects

Kaius Sinnemäki
University of Helsinki
P.O. Box 24
00014 University of Helsinki
kaius.sinnemaki@helsinki.fi

Akira Takaki
University of Helsinki
P.O. Box 24
00014 University of Helsinki
akira.takaki@helsinki.fi

Abstract

In this paper, we research the interaction of number agreement, dependency length, and word order between the subject and the verb in Finnish traditional dialects. While in standard Finnish the verb always agrees with the subject in person and number, in traditional dialects it does not always agree in number with a third person plural subject. We approach this variation with data from The Finnish Dialect Syntax Archive, focusing here on plural lexical subjects. We use generalized linear mixed effects modelling to model variation in number agreement and use as as a predictor the dependency length between the subject and the verb, building in word order as part of this measure. Variation across lemmas, individuals, and dialects is addressed via random grouping factors. Finite verb and the main lexical verb are considered as alternative reference points for dependency length and agreement. The results suggest that the probability of number agreement increases as the distance of the preverbal subject from the verb increases, but the trend is the opposite for postverbal subjects so that the probability of number agreement decreases as the distance of the subject from the verb increases.

1 Introduction

Over the past two decades dependency relations have been much researched from the perspective of dependency length. Dependency length measures the distance between the head and the dependent of a construction in terms of the number of intervening words. Cross-linguistic research suggests a tendency to keep dependency length minimal across languages (Hawkins, 2004; Liu et al., 2017; Gibson et al., 2019; Jing et al., to appear). Interaction of dependency length with other grammatical factors, such as word order, has also been increasingly researched. However, there has been very little research on the possible relationship between dependency length and variation in case marking and/or agreement (Ros et al., 2015; Sinnemäki and Haakana, 2021) despite increasing calls for doing so. Most previous research also focuses on written language or a mixture of spoken and written language using, for instance, the Universal Dependencies data (Zeman et al., 2021; de Marneffe et al., 2021).

In this paper we discuss the interaction of number agreement on the verb and the length of dependency between the lexical subject and the verb in Finnish traditional dialects, thus focusing on spoken language varieties. Verbs in standard spoken Finnish agree obligatorily with the subject in person and number, as in example (1a), so that using the singular form of the verb with plural subjects is ungrammatical in the standard language. However, third person plural subjects do not always trigger plural agreement on the verb in colloquial speech and in dialects, as in example (1b).

(1) a. *lapse-t* *syö-*ø/vät*
child-PL.NOM eat-3SG/3PL
'children are eating'

b. *lapse-t* *syö-ø/vät*
child-PL.NOM eat-3SG/3PL
'children are eating'

Previous work on this variation has suggested that plural agreement on the verb may be affected by different factors. These include sociolinguistic factors, such as speakers gender and dialect, as well as structural factors. For instance, plural agreement is rare with the copula verb, quite common with preverbal subjects, and quite likely when the subject is far removed from the verb (Karlsson, 1966; Karlsson, 1977; Mielikäinen, 1984). This earlier research thus already suggests that dependency length and word order affect plural agreement. There is also much cross-dialectal variation in number agreement, the plural agreement being the most frequent in the South-Eastern, the South-Western, and the Northernmost dialects but uncommon elsewhere. However, the relative effect of these factors have not been evaluated with one another using computational modelling, taking into account dialectal variation as well.

In this paper we focus on the interaction of number agreement, word order, and dependency length using corpus data on Finnish traditional dialects and modelling variation in agreement computationally. We are specifically interested in how word order and dependency length may affect variation in number agreement. While number agreement in Finnish varies in different constructions, we focus here on number agreement on the verb, because this variation is well-covered in earlier literature and provides an interesting foundation for further research.

We take as a starting point the noisy channel hypothesis, according to which language users are sensitive to how noise may corrupt the linguistic signal (Gibson et al., 2013). In the case of the dependency relation between the subject and the verb, one source of noise are words that intervene between the subject and the verb. The more such intervening words there are, the more this burdens the memory and may hamper the hearer's ability to recover the dependency relation. When applied to variable plural agreement, the noisy channel hypothesis predicts that the greater the distance between the plural subject and the verb, the more likely the verb will agree with the plural subject to maximize the hearer's ability to recover the dependency relation. But when the subject and verb are very close to each other, there is less noise from intervening words and thus the likelihood of plural agreement is predicted to be low. Other grammatical structures, such as repeating the verb, may be used for maximizing the recoverability of the dependency relation especially in spoken language, but these structures are excluded from this study.

These predictions are further qualified by word order. With plural preverbal lexical subjects, agreement is the only reliable source of information for the dependency relation in Finnish, since both plural lexical subject and objects may be in the nominative case. Because the order of subject and verb is very flexible in Finnish dialects (see Section 2), word order is not informative about syntactic structure either. However, the verb's argument structure may provide information about the arguments at the verb. Given these sources of information for recovering the dependency relation, we predict that plural agreement is more likely with preverbal than with postverbal subjects. This prediction accords also with what is known about plural agreement in the world's languages. Based on earlier research there is a universal tendency to suspend plural agreement between the subject and the verb in postverbal contexts (Greenberg, 1966), that is, to use singular verb forms with postverbal plural subjects. This pattern is found in standard Finnish as well (Karlsson, 1977).

We model the effect of dependency length on the variation in number agreement with generalized linear mixed effects modelling. The null hypothesis is that dependency length has no effect on number agreement. In the modelling we take into account variation in word order and address variation in number agreement across speakers, dialects, and lemmas as well. The data comes from roughly 4 500 clauses retrieved from The Finnish Dialect Syntax Archive (University of Turku, School of Languages and Translation Studies and Institute for the Languages of Finland, 1985). In the following, we first discuss the data and methods (Section 2), followed by the results of the statistical modelling (Section 3) and a brief discussion of the results (Section 4).

2 Data and methods

Based on earlier research variation in number agreement is particularly common in Finnish traditional dialects. For this reason, we analysed data from The Finnish Dialect Syntax Archive (University of Turku, School of Languages and Translation Studies and Institute for the Languages of Finland, 1985), which contains recorded spoken data from more than 100 interviewees, totaling roughly one million

lemmas.[1] The data has been collected between the 1950s and 1970s and contains largely narratives from uneducated rural residents whose speech has not been affected by the standard language (Ikola, 1985). The interviewees' median year of birth was 1884, so the data represents Finnish dialects as learned at the end of the 19th century when standard language was taking shape but had not had a widespread effect on the population. The Archive's data is grammatically annotated and contains information, for instance, on the speakers age, gender, and dialect as well as grammatical information on each word (e.g., part of speech, inflectional categories, and syntactic function).

We extracted the data using the following criteria.[2] First, we contrasted two ways of defining the head of the construction. Dependency length is analysed as the distance between the head (the verb) and the dependent (the subject). However, when the predicate is composed of several parts, each of which can agree with the subject in number, the situation becomes more complex: how should we account for number agreement on an inflecting auxiliary that is closer to the subject compared to the main lexical verb? It is plausible to assume that placing the auxiliary close to the subject would enable earlier identification of the dependency relations (Ros et al., 2015, p. 1160-1161).

In Finnish, agreement on the predicate can be expressed on three different elements. Example in Figure (1) illustrates how not only the main lexical verb (*syödä* 'to eat') can agree with the subject in number but so can the auxiliary verb (*olla* 'to be') and the negative auxiliary verb (*ei*). Such complex predicates pose a potential problem for analysing the relationship between agreement and dependency length. In this paper we contrast two ways of approaching this issue. We start by modelling the finite verb as the head, that is, as the reference point for dependency length and agreement. In the case of simple verbs the main lexical verb is also the finite verb. In the case of complex verbs, the finite auxiliary is the finite element, while the main lexical verb is non-finite. We then contrasted this approach by modelling the main lexical verb as the head. However, in the case of complex verbs with three elements, the non-finite auxiliary verb (*olla* in the example in Figure 1) could be considered as an alternative reference point for dependency length and agreement as well. This was not attempted here, since there were only 14 such instances and in each of them the auxiliary was in the singular.

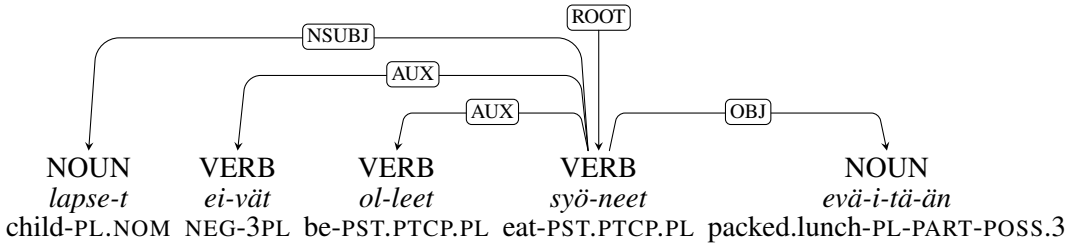

Figure 1: Dependency tree of the Finnish sentence 'The children had not eaten their packed lunches'.

Second, we limited the analysis to clauses containing a lexical subject and excluded pronoun subjects from the study. The reason for this was that earlier research on third person plural subject pronouns has already suggested that a growing distance between the subject pronoun and the main lexical verb increases the probability of plural forms on the verb at least with preverbal subjects (Sinnemäki and Haakana, 2021). There are also two third person plural pronouns in Finnish, namely *ne* and *he*. The latter pronoun is much less common across Finnish dialects but it also occurs much more frequently with plural agreement compared to *ne*. For these reasons, we thought it would be meaningful to focus only on lexical subjects and to contrast also the preverbal and the postverbal domains.

Third, while the corpus is carefully annotated for grammatical information, it does not currently code dependency relations as treebanks do. For this reason, we automatically extracted all relevant clauses and then manually double-checked each verb-subject pair for dependency length, word order, and overall

[1]The whole corpus is openly available via the Language Bank of Finland at `http://urn.fi/urn:nbn:fi:lb-2019092002`.

[2]The analysed data and the scripts are available at `https://version.helsinki.fi/gramadapt/depling2021-number-agreement`.

correctness of the analysis. In general, the greater the initial dependency length was, the more likely it was wrongly analysed by our automatic extraction. There were also some cases where the verb was repeated multiple times before the subject, which led to suspiciously long dependency lengths in the automatic analysis. In the manual analysis, the dependency length for such sentences was analysed from the nearest verb to the subject. One of the most extreme cases is illustrated in (2).

(2) *nii sitte oli tuola täälä ojala-sa justihin siittä kajuuti-lta ojalankylä sielä oli ni*
so then be.PST there here Ojala-INE right there.from Kajuuti-ABL Ojala.village there be.PST yes
oli kinkerit
be.PST reading.exams
'So, there were reading examinations at Ojala-village, right at Kajuutti.'

In this example, the distance between the first copula *oli* and its subject *kinkerit* is 12.[3] However, the copula is repeated twice before the lexical subject and the closest copula is actually adjacent to the subject. By and large the automatic analysis of dependency lengths were correct in subject-verb orders, but in verb-subject clauses about a quarter were discarded, because the verb was preceded by another subject, often an anaphoric pronoun. This was expected to some extent, as Finnish is an SVO language. Following these criteria the final data contains 4 561 clauses.

Although the annotation of the original corpus has been meticulously refined over the years, it may still contain errors. For instance, we corrected 46 lemmas (roughly 1%) that were wrongly analysed in the original. It is possible that some subject-verb dependencies were overlooked by our automatic extraction, potentially leading to some false negatives (that is, excluding instances that should have been included). However, since our extraction method relied on the annotations, the potential false positives would most likely stem from problems in the original annotation. We did not estimate the correctness of the original annotations in this regard but suspect the rate of unrecognised dependencies is very low.

Length of dependency is defined as the number of intervening words between the head and the dependent in a construction. For the purpose of modelling, we coded dependency length following Gildea and Temperley (2010) so that it received negative values in left-branching dependency-relations, that is, where the subject preceded the verb (the finite auxiliary or the main lexical verb), and positive values in right-branching dependency-relations, that is, where the subject followed the verb, the head (the verb) itself at zero. This coding enables us to keep the ensuing model structure simple and to put emphasis on dependency length in the modelling, while still being able to inspect linear order at least visually. An alternative would have been to use positive counts for dependency length and to model its interaction with word order. Because this would have increased the complexity of the model we opted for coding dependency length with both positive and negative values.

Figure 2 displays the histogram for dependency length over agreement. In both plots, the majority of instances is adjacent to the verb with diminishing number of instances as the distance from the verb grows. The subject tends to occur mostly preverbally, but postverbal lexical subjects are also common with both finite verbs (plot A) and main lexical verbs (plot B). Overall, there is a lot of variation in the order of the subject and the verb in Finnish dialects. The distribution of number agreement is biased so that plural forms of the verb are relatively more common among preverbal lexical subjects, while singular forms are relatively more common among postverbal lexical subjects. In addition, plural agreement seems slightly more common as the preverbal subject is further removed from the verb and singular agreement seems slightly more common as the postverbal subject is further removed from the verb. Yet based on the histograms alone it is hard to draw conclusions on how number agreement behaves more generally as distance from the verb increases.

To estimate whether dependency length has an effect on number agreement, we used mixed effects logistic regression. Number agreement was modelled as a binomial response variable with values "singular" (reference level) and "plural". Dependency length between the subject and the verb was modelled as a predictor, counted as the number of intervening words as stated above. Two different models were contrasted. In the first model, called here m.fin, the finite auxiliary was analysed as the head. In these

[3] *Kinkerit* refers to examinations in rural areas held historically to teach and test reading skills and knowledge of Christianity.

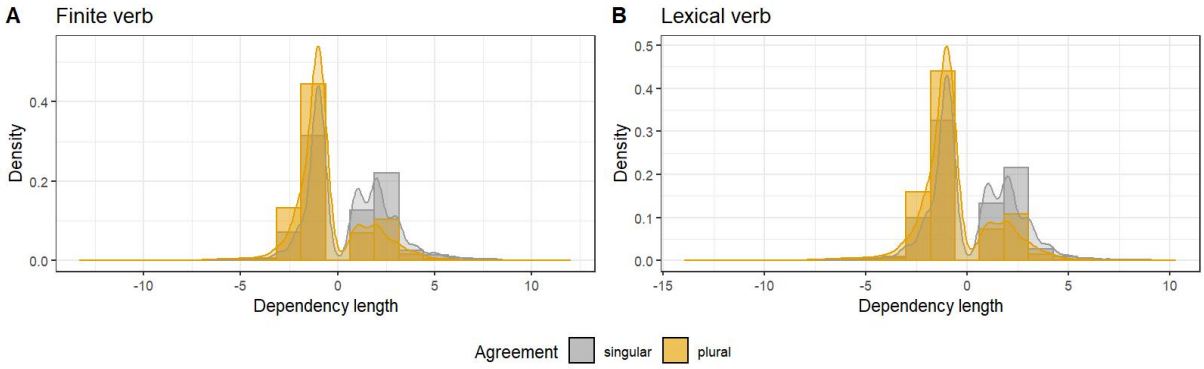

Figure 2: Histograms for dependency length over number agreement (finite verbs as the reference point in plot A and main lexical verbs in plot B).

models we analysed the occurrence of number agreement on the finite auxiliary and used it also as a reference point for counting dependency length. In the m.fin models there were 934 (21%) clauses with plural agreement; dependency length ranged from values -13 to +11, the verb being at zero. In the second model, called here m.lex, the main lexical verb was analysed as the head. In these models we analysed the occurrence of number agreement on the main lexical verb and used it also as a reference point for counting dependency length. In the m.lex models there were 1003 (23%) clauses with plural agreement; dependency length ranged from values -13 to +10, the verb being at zero.

Three random intercepts were included in both models: i. the lemma of the (main lexical) verb, ii. the lemma of the lexical subject, and iii. the individual speaker nested in their local dialect group. Based on earlier research the lemma of the verb may affect number agreement in Finnish dialects: plural agreement is particularly rare with the copula *olla*, but there is great variation across different verbs. In m.fin models there can be only two alternative finite elements, namely, the negative auxiliary *ei* or the verb *olla* which functions as an auxiliary in the perfect and pluperfect tenses. For this reason we modelled the main lexical verb as a random intercept also in the m.fin model. We also assume that variation depending on the subject lemma needs to be accounted in the modelling, analogously to the verb lemma. The hierarchic structure of embedding each speaker in their dialect group enables taking into account variation in number agreement within and across dialects and speakers.

The models were fitted in R using the package `blme` (Chung et al., 2013), which enables maximum penalized likelihood with weakly informative priors and posterior modes for estimation. It often leads to better convergence compared to `lme4` as well as drawing correlation terms away from perfect correlation. The model specification in the `lme4` notation (Bates et al., 2015) was as in (3). The p-values were drawn with likelihood ratio. The models' explanatory power was computed separately for the whole model (conditional R^2) and just for the fixed effects (marginal R^2) via the package `MuMIn` (Barton, 2020). The algorithm is based on Nakagawa and Schielzeth (2013) and has been further developed by Johnson (2014), and Nakagawa et al. (2017).[4]

(3) agreement $\sim$ dep.length + (1|lemma.noun) + (1|lemma.verb) + (1|dialect/individual)

3 Results

According to the results, dependency length had a significant negative effect on plural agreement when finite verbs were selected as the reference point ($estimate = -0.28 \pm 0.03; \chi^2(1) = 97.2; p < 0.001$). This means that as dependency length increases by one unit, the likelihood of plural agreement on the finite verb decreases about 1.25 times. When selecting the main lexical verb as the reference point, dependency length had also a significant negative effect on plural agreement ($estimate = -0.18 \pm$

[4]The R package `tidyverse` (Wickham et al., 2019) was used in preprocessing the data in R; graphics were computed using packages `sjPlot` (Lüdecke, 2020), `cowplot` (Wilke, 2020), and `ggplot2` (Wickham, 2016).

$0.03; \chi^2(1) = 39.7; p < 0.001$). This means that as dependency length increases by one unit, the likelihood of plural agreement on the main lexical verb decreases about 1.17 times.

We evaluated the models' goodness-of-fit with Akaike Information Criterion (AIC) by comparing the difference in the nested models' values for AIC. Adding dependency length to the null model m.fin lowers AIC by 95, while adding dependency length to the null model m.lex lowers AIC by 38. This large reductions in AIC (> 10) provide evidence for both models' goodness (Burnham and Anderson, 2002, p. 70-71). The explanatory power of dependency length in model m.fin was about 0.030 (marginal R^2) and for the whole model about 0.494 (conditional R^2); for model m.lex the respective figures were 0.013 and 0.500. Accordingly, most of the variation in plural agreement was explained by dialectal and individual differences, but even so the models were able to recognize a small effect for dependency length.

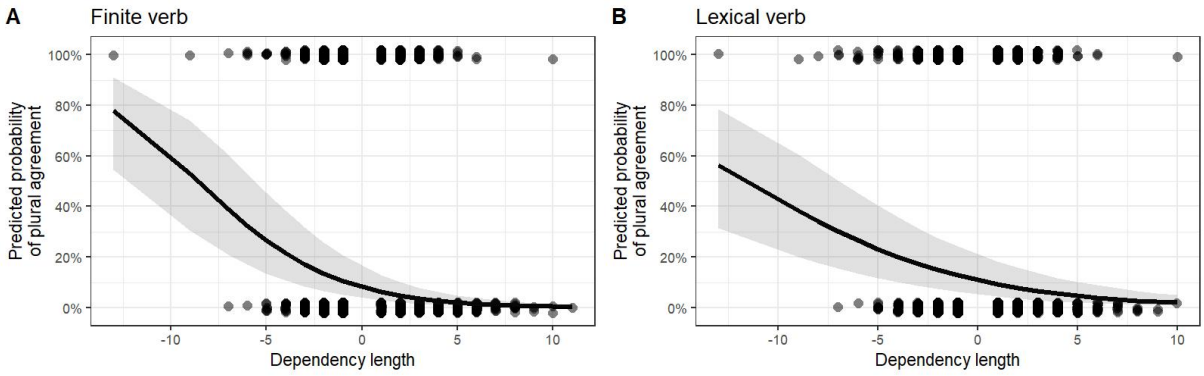

Figure 3: Marginal effects for dependency length over number agreement (in plot A for finite verbs and in plot B for main lexical verbs; small jitter is added to the datapoints).

Figure 3 presents the marginal effect plots for the two models. The plots suggest a clear inverse relationship between dependency length and number agreement. In both plots the predicted probability of plural agreement is about 10% when the plural lexical subject is adjacent to the verb. However, the more words intervene between a *preverbal* subject and the verb, the greater the predicted probability of plural agreement becomes. In plot A it is around 40% at a distance of seven and increases above 60% at the greatest distances, while in plot B it is around 30% at a distance of seven and increases above 40% at the greatest distances. On the other hand, the more words intervene between a *postverbal* lexical subject and the verb, the smaller and ever closer to zero the predicted probability of plural agreement becomes in both plots. Word order thus seems to condition the effect of dependency length on number agreement: plural agreement is more likely when the lexical subject precedes the verb than when it follows the verb, and the difference between the word orders becomes the clearer the greater the dependency length is.

4 Discussion

Based on our analyses, there was an inverse relationship between number agreement and dependency length in Finnish traditional dialects partly conditioned by word order. The inverse relationship was a little stronger with finite verbs than with main lexical verbs. But regardless of which was taken as the reference point for agreement and dependency length, the results were significant and very similar.

Since our models were random intercept models we could not estimate whether dependency length had a similar effect on agreement across dialects. To evaluate this, we fitted two further models. These models were otherwise identical to the random intercept models, but we fitted a random slope for dependency length over dialect groups (and over individuals). Because plural agreement is very unevenly distributed across dialects, we included data from only those dialect groups in which there were 20 or more instances of plural agreement and where that incidence was 10% or more of all the instances.

According to the results, dependency length had a significant negative effect on plural agreement with finite verbs ($estimate = -0.37 \pm 0.08; \chi^2(1) = 13.0; p < 0.001$) as well as with main lexical verbs ($estimate = -0.32 \pm 0.09; \chi^2(1) = 10.4; p = 0.0013$). The marginal effects in Figure 4 are

quite similar across the dialects regardless of using finite verbs (plot A) or main lexical verbs (plot B) as reference points for dependency length and number agreement: the farther a *preverbal* lexical subject is removed from the verb, the *more likely* there is plural agreement on the verb, and the farther a *postverbal* lexical subject is removed from the verb, the *less likely* there is plural agreement on the verb. These results suggest the relationship between agreement and dependency length is similar across the traditional Finnish dialects and regardless of which verb was selected as the reference point.

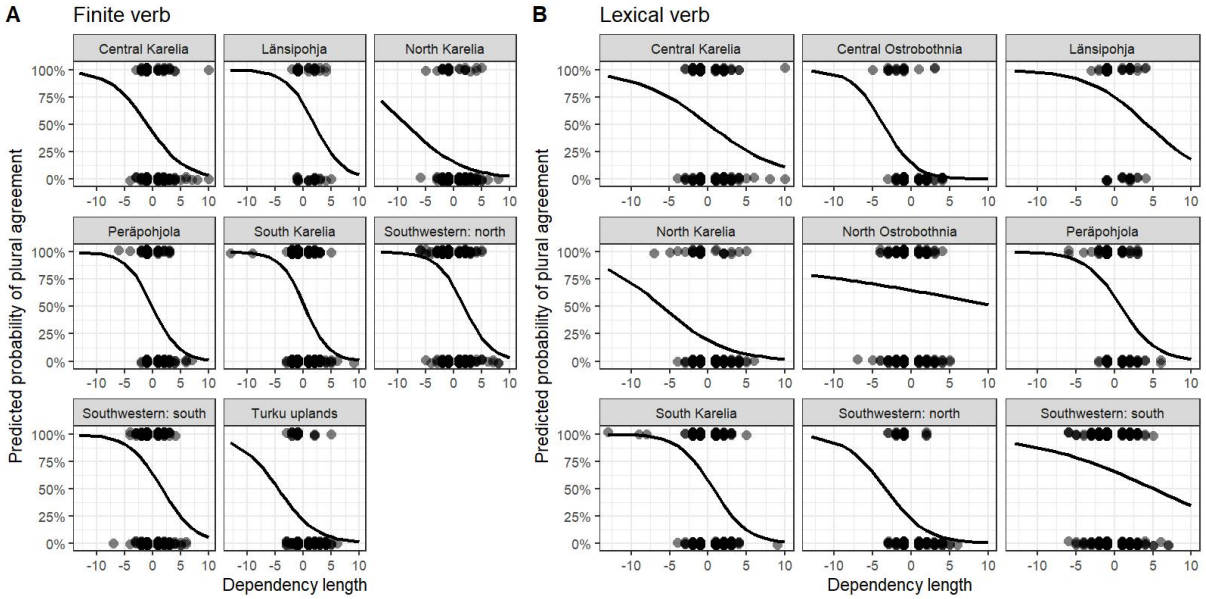

Figure 4: Marginal effects for dependency length over number agreement in the random slope models.

The results largely support our predictions based on the noisy channel hypothesis. Plural agreement increased in probability as more words intervened between the subject and the verb. This result aligns with earlier research on third person plural pronoun subjects in Finnish (Sinnemäki and Haakana, 2021). We also predicted that plural agreement would be less likely with postverbal subjects compared to preverbal subjects, and the results provide evidence for this hypothesis as well.

However, it was somewhat unexpected that the probability of plural agreement became increasingly smaller the farther the postverbal lexical subject was removed from the verb. While the results align with how other languages work (Greenberg, 1966), it is unclear why plural agreement would be less likely with postverbal subjects far removed from the verb compared to postverbal subjects that were adjacent to the verb. In the postverbal contexts in Finnish, the subject may be more easily confused with the object, because direct objects tend to occur postverbally and since plural lexical objects as well as plural lexical subjects may occur in the nominative case (objects also in the partitive case). It would thus seem that there were more possibilities for confusing the subject and the object in the postverbal domain, which, according to the noisy channel hypothesis, would call for increased probability of agreement with postverbal subjects, at least for transitive and ditransitive verbs. Further research is needed to determine which factors affect variation in plural agreement especially in the postverbal domain.

The results raise a more general question whether the observed relationship between number agreement and dependency length is limited to Finnish dialects or a more general tendency in languages. We do not consider it implausible that number agreement and dependency length would pattern in similar ways in other languages as well, but this remains as an issue for future research, since the interaction between dependency length and agreement has not yet been widely researched across languages.

Acknowledgements

This research has received funding by the European Research Council (ERC), grant no 805371 to Kaius Sinnemäki (PI). We are grateful to three anonymous reviewers and to the editors for comments.

References

Kamil Barton. 2020. Mumin: Multi-model inference. r package version 1.43.17.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Kenneth P. Burnham and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, second edition.

Yeojin Chung, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. 2013. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4):685–709.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Joseph H. Greenberg. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA, second edition.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford.

Osmo Ikola, editor. 1985. *Lauseopin arkiston opas*, volume 1 of *Lauseopin arkiston julkaisuja*. Turun yliopisto, Turku.

Yingqi Jing, Damin E. Blasi, and Balthasar Bickel. to appear. Dependency length minimization and its limits: a possible role for a probabilistic version of the final-over-final condition. *Language*.

Paul C.D. Johnson. 2014. Extension of nakagawa & schielzeth's r2glmm to random slopes models. *Methods in Ecology and Evolution*, 5(9):944–946.

Göran Karlsson. 1966. Eräitä tilastollisia tietoja subjektin ja predikaatin numeruskongruenssista suomen murteissa. *Sananjalka*, 8:2–23.

Fred Karlsson. 1977. Syntaktisten kongruenssijärjestelmien luonteesta ja funktioista. *Virittäjä*, 81(4):359–391.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Daniel Lüdecke, 2020. *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.4.

Aila Mielikäinen. 1984. Monikon 3. persoonan kongruenssista puhekielessä. *Virittäjä*, 88(2):162–175.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Shinichi Nakagawa, Paul C.D. Johnson, and Holger Schielzeth. 2017. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213.

Idoia Ros, Mikel Santesteban, Kumiko Fukumora, and Itziar Laka. 2015. Aiming at shorter dependencies: The role of agreement morphology. *Language, Cognition and Neuroscience*, 30(9):1156–1174.

Kaius Sinnemäki and Viljami Haakana. 2021. Variationistinen korpustutkimus predikaatin differentiaalisesta lukukongruenssista ja substantiiviluokasta suomen murteissa. In Leena Maria Heikkola, Geda Paulsen, Katarzyna Wojciechowicz, and Jutta Rosenberg, editors, *Språkets funktion: Juhlakirja Urpo Nikanteen 60-vuotispäivän kunniaksi-Festskrift till Urpo Nikanne på 60-årsdagen-Festschrift for Urpo Nikanne in honor of his 60th birthday*, pages 96–130. Åbo Akademis förlag, Åbo.

University of Turku, School of Languages and Translation Studies and Institute for the Languages of Finland. 1985. The Finnish Dialect Syntax Archive's Helsinki Korp Version.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Franois, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Mller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Claus O. Wilke, 2020. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.1.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.