A Multi-Model Analysis of Variation: How many subjunctives are there in Spanish?

Introduction: Spanish has three simple morphologically distinct subjunctive forms: two past tenses *-se* and *-ra* and one present tense. However, it is well known that the separation of labor between present and past subjunctive is not always clear. Many dialects have been reported to use present tense under past tense main predicates in unambiguously past contexts (Sessarego 2008, 2010, Del Rio 2014, Guajardo and Goodall, to appear). On the other hand, it is not clear whether there is any semantic distinction between the two morphologically past forms (Guzmán Naranjo 2017, Rosemeyer and Schwenter 2019). Thus the question remains how many distinctions are necessary to account for the variation in simple subjunctive forms?

Research Questions: (1) What type of model makes a more accurate classification of subjunctive forms in past obligatory subjunctive contexts: a model with three separate forms (*-se*, *-ra* and present) or a model with two forms (*present* vs. *past*)? (2) What are the most important variables driving the variation in subjunctive forms in Spanish and what is the directionality of their effect?

Methodology: I address these questions by conducting a corpus study using the Web/Dialect version of Corpus del Español (Davies 2002) coupled with random forest and logistic regression modeling. Data was extracted for six obligatory subjunctive main predicates in the past (preterite and imperfect) with the three subjunctive forms in three dialects: Argentinean, Mexican and Peninsular Spanish. A stratified random sample technique was used to create a dataset for each country with 600 sentences each. Because the data was very imbalanced with respect to *-se* and *present* forms *(-ra* was overrepresented), I used AdaSyn (Haibo *et al* 2008) to create a balanced dataset so that every class contained a sufficient number of tokens for the classification model (i.e., the random forest) to learn from. AdaSyn is an adaptive synthetic approach for learning from imbalanced datasets. By calculating a weighted distribution for the minority class examples using k-nearest neighbors, it generates new synthetic examples of the minority classes that are harder to learn. The final dataset contained 3859 sentences (34.5% *present*, 33.4% *-ra, 32% -se)*. Two types of analyses were conducted with each methodology: a multiclass analysis with the three forms as separate outcomes of the dependent variable SUBJUNCTIVE TYPE *(-se, -ra, present)* and a binary analysis where the two past subjunctives were collapsed as past tense, with the dependent variable being SUBJUNCTIVE TENSE (*pres.* vs *past*). R 3.6.1 was used for all statistical modeling (R core Team 2019). For the classification modeling, the data was analyzed using random forests (Breiman 2001) in the "party" package (Hothorn et al 2006, Zeilis et al 2008, Strobl et al 2007, Strobl 2008). The logistic regressions were conducted with the "nnet" package for the multinomial regression analysis (Venables and Ripley 2002) and "lme4" for binary logistic regression (Bates *et al* 2015). The original datasets of 600 sentences per country (1798 total) were used for the regression analysis. Each random forest was built with 2001 trees and conditional variable importance was calculated for each analysis. The variable importance results were then used to build the logistic regression models in order to establish the way in which each variable contributed to the variation.

Results: The multiclass random forest analysis yielded an Accuracy score of 0.78 (95% CI: 0.77, 0.80) with an area under the curve (AUC) score of 0.86. The conditional variable importance yielded seven variables as the most important ones: SOURCE (proxy for register), COUNTRY, FREQUENCY OF THE VERB IN THE SUBJUNCTIVE, NUMBER OF MAIN PREDICATE, NUMBER OF SUBJUNCTIVE VERB, MAIN VERB AND TENSE OF MAIN PREDICATE. The variables with **no predictive power** were: VERB CLASS, GRAMMATICAL CATEGORY AND DEFINITENESS OF THE SUBJUNCTIVE VERB SUBJECT, THE LEXICAL VERB IN THE SUBJUNCTIVE, GRAMMATICAL

CATEGORY AND DEFINITENESS OF THE DIRECT OBJECT, PERSON AND ANIMACY OF THE SUBJUNCTIVE VERB SUBJECT AND GRAMMATICAL ASPECT OF THE SUBJUNCTIVE VERB.
The binary random forest analysis yielded a much higher Accuracy score of 0.89 (95% CI: 0.88, 0.90) with an AUC score of 0.88. The conditional variable importance ranking was: COUNTRY, SOURCE, NUMBER OF MAIN PREDICATE, NUMBER OF SUBJUNCTIVE VERB, FREQUENCY OF THE VERB IN THE SUBJUNCTIVE and MAIN PREDICATE. The same variables deemed unimportant in the multiclass analysis had no predictive power in the binary analysis. The multinomial regression (McFadden R^2= 0.20, Accuracy = 0.73) determined that *-se* is disfavored with verbs that take infinitival complements ($p < 0.01$) and when the main verb appears in the preterite ($p < 0.05$) but it is more likely to appear in more formal registers ($p < 0.01$). The present tense appears significantly less with telic predicates ($p <0.01$) and with singular main verbs ($p < 0.01$) and singular embedded verbs ($p < 0.01$). Relative to Spain, *-se* is significantly less likely to appear in both Argentina and Mexico ($p < 0.01$ for both) but no significant difference was found between Mexico and Spain in the use of present tense, while in Argentina the chances of the present tense in past contexts are almost 20 times greater than in Spain. The binary logistic regression goodness-of-fit measures were much higher than the multinomial model (McFadden R^2= 0.28, AUC=0.87) suggesting a better performance and fit of the model. Again, there was no significant differences between Mexico and Spain ($p = 0.69$) but in Argentina the chances of finding a present tense in past contexts is 13 times higher than in Spain ($p < 0.001$) and the present tense is disfavored in Spain when the subjunctive verb is singular ($p < 0.005$). There were two types of interactions between country and number of the subjunctive verb and between the number of the main verb and the main verb itself. The first interaction is driven by the fact that in Mexico and Argentina the present tense is favored with singular verbs (unlike in Spain). The second interaction shows that when the predicate *querer* appears in the singular, it disfavors the present tense ($p < 0.005$).

Discussion: Both the random forest and the logistic regression models achieved much higher goodness-of-fit parameters with the grammar with a binary split (*pres* vs. *past)*. This indicates that no loss of information is incurred when the two past subjunctive classes are collapsed into one single class, suggesting that in complement clauses to obligatory subjunctive predicates there appears to be no semantic difference between these two morphologically distinct subjunctive forms. The variables with the most predictive power in the two analyses were similar but the ranking was slightly different. However, the tense of the main verb appears as an important variable in the multiclass random forest but not in the binary split, and we saw in the logistic regression that this is because the form with *-se* is dispreferred when the main verb is in the preterite. These results provide quantitatively robust evidence that the two past subjunctive forms, at least in complement clauses to obligatory subjunctive predicates, do not differ in a semantically meaningful way. A grammar that only posits a *present* vs. *past* opposition explains the data more accurately. This paper also highlights how two complementary methodologies as random forests and logistic regression can be used together to arrive at a more comprehensive and deeper understanding of linguistic variation.

Selected references:

Guzmán Naranjo, M. (2016). The se-ra Alternation in Spanish Subjunctive. *Corpus Linguistics and Linguistic Theory*, 13(1), pp. 97-134.

Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 1322-1328.

Rosemeyer, M. & Schwenter, S. (2019). Entrenchment and persistence in language change: the Spanish past subjunctive . *Corpus Linguistics and Linguistic Theory*, 15(1), pp. 167-204.