

Author Query Form

The se-ra Alternation in Spanish Subjunctive

Article: CLLT-2015-0017

Query No	Page No	Query
Q1	3	Please check that the short title is OK.
Q2	19	Please check the usage of the word random forest and Random forest, capitalization is inconsistently given. Please clarify which one is correct.
Q3	35	Please provide Publisher Location for reference “Aronoff 1995”.
Q4	35	Please provide editor name, Publisher Location for reference “Baayen 1992”.
Q5	36	Please provide editor name for reference “Bresnan et al. 2007”.
Q6	36	Please provide Publisher Location for reference “Cuervo et al. 1981”.
Q7	36	Please provide Publisher Location for reference “Gili Gaya 1983”.
Q8	36	Please provide Publisher Location for reference “Gries 2003”.
Q9	37	Please provide VolumeID for reference “Kempas 2011”.
Q10	37	Please provide PublisherLocation for reference “Rojo 2008”.
Q11	37	Please provide PublisherLocation for reference “Steels 2011”.

Matías Guzmán Naranjo*

1

The se-ra Alternation in Spanish Subjunctive

DOI 10.1515/cllt-2015-0017

5

Abstract: In this paper I take a look at a classic problem in Spanish morphosyntax, namely the alternation between the forms *-se* and *-ra* in the Imperfect Subjunctive (*Imperfecto de Subjuntivo*). Research on this topic has mainly focused on sociolinguistic variation, and has been done almost exclusively with impressionistic data and speakers' intuitions. I address the problem from a usage-based perspective, using corpus linguistics methods. The main claim is that the choice between *-se* and *-ra* correlates to a certain extent with morpho-syntactic and discourse factors. Through collocation analysis I also show that there exists repelled and attracted collexemes that distinguish and relate both forms.

15

Keywords: construction morphology, naive discriminative learning, *-se/-ra* Spanish alternation

1 Introduction

20

The morphological alternation between *-se* and *-ra* in the Spanish *imperfecto del subjuntivo* ("imperfect subjunctive") has been studied extensively but it is still poorly understood, and remains a challenging problem. The alternation is shown in (1).

25

- (1) a. Si yo fuera ingeniero no estaría en esta
if I be.1SG.IMP.SBJ engineer no be.1 SG.COND.PRES in this
situación.
situation
"If I were an engineer I wouldn't be in this situation"
- b. Si yo fuse ingeniero no estaría en esta
if I be.1SG.IMP.SBJ engineer no be.1SG.COND.PRES in this
situación.
situation
"If I were an engineer I wouldn't be in this situation"

30

35

*Corresponding author: Matías Guzmán Naranjo, Institut für Linguistik, Universität Leipzig, Beethovenstraße 15 D-04107, Leipzig, Sachsen, Germany,
E-mail: matias.guzman_naranjo@uni-leipzig.de

40

Both forms are, at least in principle, possible with all Spanish verbs, and there is no categorical distinction in their use. The difference between the two is elusive and hard to pin down. Most research on this alternation has so far tried to characterize its sociolinguistic aspects focusing mainly on how dialects differ in the attested proportions of use (see Section 2), but little is known regarding its distributional properties within dialects, and even less is known about how and why speakers choose one form or the other.

This paper deals exclusively with the latter, that is, what factors are correlated with the use of *-se* or *-ra*, what patterns are present in corpora, and how predictable the alternation is from the morpho-syntactic and discourse context. I will deal exclusively with Peninsular Spanish and will ignore issues related to dialectal variation (for some discussion of sociolinguistic variation and dialectal aspects of this alternations see for example (Rojo 2008; Kempas 2011) and references therein).

The structure of the paper is as follows. Section 2 briefly discusses some of the previous work that has addressed the *-se/-ra* alternation, and tries to characterize the types of methods that have been used so far. Section 3 sketches a simple constructional analysis of the alternation based on work by (Booij 2010a), which will be used as a starting point for the empirical investigation. Section 4 describes the materials and methodology used for this study. Section 5 presents a brief discussion on the relative productivity of both forms. Section 6 describes the distribution of *-se* and *-ra* in the corpus studied. Section 7 presents a Naive Discriminative Learning model that shows how different morpho-syntactic and discourse properties of the context correlate with *-se* and *-ra*. Section 8 reports on a collostructional analysis for both forms, and what collexemes can tell us about the semantics of the construction. Section 9 provides some discussion of the results and section 10 offers some final remarks.

All statistical tests, plots and models were done using R programming language (R Core Team 2014).

30

2 Previous work on the *-se/-ra* alternation

There has been extensive research into the Spanish imperfect subjunctive for the last hundred and forty years or so, but it has overwhelmingly focused on interspeaker variation, and on dialectal differences that exists between Spanish speaking communities. In this section I very briefly summarize some of the most prominent investigations on the matter and their overall conclusions (for a more comprehensive discussion see DeMello (1993), for example).

40



The form *-se* evolved from the Latin plusquamperfect subjunctive, while the form *-ra* evolved from the Latin plusquamperfect indicative (Wilson 1983). According to Cuervo and Ahumada (1981) the form *-ra* started to be associated with an indicative mood and slowly acquired the subjunctive mood over time through analogy with the form *-se*. Today *-se* and *-ra* are seen as two near synonymous morphemes in free variation. As early as 1874 Cuervo and Ahumada (1981) note that there was a significant difference in the proportion of both forms between American Spanish and Peninsular Spanish. Although they do not give numbers, they claim that Spaniards used *-se* almost exclusively, and that this form was almost absent in casual speech in America. Cuervo and Ahumada also claim that *-se* was used in Colombia mainly by writers who were trying to imitate peninsular varieties.

Wilson (1983) traces the evolution of *-se* and *-ra* in the Mexican written language, but treats both forms as having converged into having an identical function. He claims that originally *-se* was the most common form used by the Conquistadores in Mexico, but that its use has steadily declined to a point of being almost nonexistent, while the use of *-ra* has become widespread.

Gili Gaya (1983, 180–181) observes regional and personal preferences in the use of *-ra* and *-se*. Additionally he claims that the form *-ra* is less frequent than the form *-se* in ordinary conversation in Spain, but that *-ra* is also in use in the written form and among educated speakers. He also claims (citing Lenz (1920)) that when one of the two forms is predominant in a dialect, then the other form is seen as more formal or pertaining to literary style.

DeMello (1993) looks at the use of both forms in Bogota, Buenos Aires, Caracas, Havana, Lima, Madrid, Mexico City, San Juan (Puerto Rico), Santiago (Chile) and Seville. His research shows that there is considerable dialectal variation between these areas, and that the proportions of both *-se* and *-ra*, as well as their functions (subjunctive or replacing the conditional) are quite different from city to city. His work, however, only focuses on dialectal variation and does not look into intra-speaker variation. His main conclusion is that although *-se* is considerably less frequent than *-ra*, the former can still be found in Spain and America, and is by no means dead.

DeMello also talks about the indicative use of the imperfect subjunctive (*el equipo que perdie-ra/se el día de ayer* “The team that lost. IMP.SUBJ yesterday”). He argues that already around 1950 its use was stilted and only present in pedantic writers. The only exceptions seem to be Argentinian Spanish, where it still seems to be fairly common, and Chilean and Cuban Spanish, where it is occasionally found.

The studies mentioned above, with the exception of DeMello’s, are all done with impressionistic data, and most of them rely solely on the author’s intuition

of what the distribution of the forms is. DeMello introduces the use of corpora to study the alternation, but he does not make use of advanced quantitative techniques, and limits himself to looking at raw frequencies.

To my knowledge, there are only two studies dealing with the *-se/-ra* alternation that make use of quantitative corpus linguistics methods. These are Schwenter (2013) and Elias and Mojedano (2014). Schwenter looked at a large number of examples¹ from different countries in the CREA corpus (Academia Española 2011) and fit a mixed effect logistic regression model to the data. In his presentation Schwenter claims to have found priming effects: when a speaker uses *-se*, he is more likely to use *-se* again when producing another imperfect subjunctive form shortly after the previous one. He also finds some effects of PERSON and NUMBER on the choice of the morpheme. However, Schwenter does not provide in his slides any accuracy scores or any other metric that allows evaluation of the model. This means that we do not know how his model performs and how many cases it can correctly predict. It is therefore not possible to contrast his results with those of the present study in any meaningful way. Elias and Mojedano (2014) report on a corpus study that looked at the historical development of the *-se/-ra* alternation using a method similar to that of Schwenter's. They claim to have found changes in how the predictors correlate with the alternation. However, because their results are not yet public we cannot know exactly what they found.

In summary, most studies on the *-se/-ra* alternation have been carried out without the use of quantitative corpus linguistic methods, and although it is well understood what the origins of this alternation are, we still know very little about its current usage in terms of its statistical and distributional properties, as well as the factors that influence the choice between the two forms.

3 The imperfect subjunctive construction

There are many options for analyzing the *-se/-ra* alternation. One natural possibility is to assume that *-se* is an allomorph of *-ra* which can be chosen freely by speakers. This seems to be the usual assumption, although it has never been articulated as such. Another option is to view both forms as different, near

¹ However, an important shortcoming of Schwenter's study is that he only considers 15 different verb types. This was presumably done so for practical reasons, but as we will see in the following sections the variable *VERB* plays the most interesting role in the *-se/-ra* alternation.

synonymous, morphemes. Both explanations are problematic. Considering *-se* 1 and *-ra* as allomorphs does not explain their systematic differences, and considering them as different morphemes does not explain their similarities and identical grammatical function.

In this paper I take a constructional view, which could be seen as a middle 5 way between the two alternatives. Following the notation proposed by Booij (Booij 2010a, 2010b, 2013) I will take the construction for the imperfect subjunctive to be as in (2).²

$$(2) \quad [[X_{vi}] - Y_{(se/ra)}]_v \leftrightarrow [SEM_i \text{ in imperfect tense subjunctive} + PRAG_1] \quad 10$$

What (2) says is basically that there is a semi-abstract imperfect subjunctive construction which combines with a verbal lexical construction X_i with a morpheme slot Y which can be either *-se* or *-ra* (but is still not specified), and produces a conjugated verb in the imperfect subjunctive associated with some pragmatic value³ not derivable from either the morpheme nor the verb. In this 15 analysis both *-se* and *-ra* are more specific constructions that instantiate the more general abstract construction in (2) and have the forms in

$$(3) \quad \begin{array}{ll} \text{a. } [[X_{vi}] - ra_j]_v \leftrightarrow [SEM_i \text{ in imperfect tense subjunctive} + PRAG_1 + PRAG_j] & 20 \\ \text{b. } [[X_{vi}] - se_k] \leftrightarrow [SEM_i \text{ in imperfect tense subjunctive} + PRAG_1 + PRAG_k] \end{array}$$

What the analysis in (3) mean is that both constructions *-se* and *-ra* instantiate the same grammatical core construction in (2) and retain the pragmatic value associated with it ($PRAG_1$) but specify additional pragmatic information exclusively associated 25 with the specific form in question ($PRAG_j$ and $PRAG_k$). This analysis correctly captures the fact that both constructions have indeed the same grammatical function, but that there seem to be important differences between the two forms. The interesting issue thus is to investigate what $PRAG_j$, $PRAG_j$ and $PRAG_k$ actually represent.

The null hypothesis that we will test is that there is no motivation for the distribution of both forms, and that the alternation is in truly free variation. The 30 alternative hypothesis is that the choice of these forms is at least partially dependent on other variables.

2 This is a simplified version. The full system would have more constructions at more abstract 35 levels that deal independently with TAM and person and number. The representation in (2) assumes that tense, aspect and mood constructions have already been merged or instantiated.

3 Here *pragmatic* is used in a very loose sense. I take it to be any meaning that is not related to the truth semantics of the construction. In addition, it includes any usage preferences, and statistical properties of the construction. It is more related to the concept of Cognitive Models in 40 (Evans 2009, 2010).

Analyzing inflectional morphology from a construction grammar perspective is, as far as I am aware, not common practice. A notable exception is Beuls (2012), where she develops a full implementation of Spanish inflectional morphology within the framework of Fluid Construction Grammar (Steels (2011), see also Schneider (2010), and Booij (2010a) for the principles behind construction morphology). She does not address this particular alternation, however.

4 Material

10

The corpus used for this study was the Corpus Oral de Referencia de la Lengua Española Contemporánea, CORLEC (Marcos Marín et al. 1992). The COR-LEC has approximately 1,100,000 words, covers a wide range of genres and was compiled with the aim of building a representative corpus of spoken standard Peninsular Spanish. I performed some semi-automatic and manual fixes of some unicode characters, formatting errors, and tagging issues, and afterwards carried out the POS tagging with the library FreeLing (Padró and Stanilovsky 2012) using its python API.

Sentence segmentation of speech data is extremely difficult, so I decided to divide the text according to single punctuation marks, namely “.” or “:” between two words, independently of whether lower or upper case followed. Other punctuation elements like “..” or “...” were ignored and not taken to be sentence boundaries because these are used throughout the corpus to denote vacillation and small pauses made by the speakers. This procedure resulted in a division that corresponds to what the transcriber of the corpus thought was a complete utterance by the speaker, which means that some text units can be larger than sentences. This also means that some sentences contain two occurrences of the imperfect subjunctive with either identical or different forms. For the collostructional analysis all sentences were considered, but for the regression models only 200 sentences for *-ra* were randomly extracted. From this latter set of sentences (plus all the occurrences of *-se*) some cases were removed if they were clear errors or instances of a different genre, e. g. citations or people reading. After all these fixes, the total number of sentence was 184 for *-ra* and 183 for *-se*.

Besides the study by Schwenter, there are no proposals in the literature for any particular set of variables that could influence the *-se/-ra* alternation. Because of this, I also included in this study, besides the variables mentioned by Schwenter, variables that have been found to be relevant in distinguishing other alternations, irrespective of whether it seemed reasonable to include them for analyzing a morphological alternation. The variables can be divided in two

groups: those pertaining to the verb, and those pertaining to the context. 1
The variables related to the verb that appears in the imperfect subjunctive
form are presented below.

The variable *VERB* is simply which verb was used in the imperfect subjunctive, which should tell us whether there are lexical preference in the alternation. 5
Directly related variables, and suggested by Schwenter,⁴ are the following:
PERSON, *NUMBER*, *TYPE*, and *MODAL*. The variables *PERSON* and *NUMBER* are both person
and number of the verb in the imperfect subjunctive. *TYPE* is the verb ending
(often referred to in the literature as thematic vowel of the verb) *-ar*, *-er* or *-ir*.
This variable could be important for priming reasons (the vowel /a/ could prime 10
-ra and /e/ could prime *-se*). Finally, *MODAL* indicates whether the verb
appearing in imperfect subjunctive has a modal meaning. The status of modals
in Spanish is not without debate. For reasons of simplicity I took the verbs
querer “want”, *poder* “can”, *deber* “must”, *soler* “do often”, *tener* “have (to)” to
be modals. The main reasons for considering these verbs as modals is that they 15
either mostly occur with other verbs (*quiero ir a comer* “I want to go to eat”), or
because they are grammaticalizing into periphrastic constructions (*tengo que ir*
“I have to go”). As we will see in Section 8 this decision seems to be justified. We
have thus lexical variables associated with the choice of verb (*VERB*, *MODAL*, *TYPE*),
and grammatical variables (*NUMBER* and *PERSON*). 20

The second set of variables pertains to the grammatical and discourse
context that the verb appears in. I coded all *-se* sentences and the randomly
chosen *-ra* sentences for:⁵ *ANIMACY OF THE SUBJECT* (NP, pronoun, drop, null, etc.),
DEFINITENESS OF THE SUBJECT,⁶ *REALIZATION OF THE SUBJECT*, *ANIMACY OF THE OBJECT*,⁷
DEFINITENESS OF THE OBJECT, *REALIZATION OF THE OBJECT* (NP, PP, pronoun, null, etc.), 25
and *SENTENCE TYPE*. For this final variable the following types were considered:
conditional (expressing a condition on which something happens), final (expressing
desire or determination that something happen), indicative (indicative use
of the subjunctive), temporal (expressing temporal relations), adversative (com-
parison or opposition to something), and potential (other uses where possibility 30
or probability are conveyed by the subjunctive; this level contains mostly the

⁴ Schwenter’s proposal to consider whether there was priming between two consecutive forms
is not practical for the present corpus because there are not enough consecutive cases of
imperfect subjunctive. 35

⁵ See the appendix for all levels of these variables.

⁶ The value abstract for definiteness is reserved for non NP subjects and objects, and is not
related to the concept of abstract nouns.

⁷ For the category of *object* I also considered adjectival and adverbial complements when there
was no direct or indirect object to the verb. *OBJECT* could be seen here as the first postverbal
complement of the verb. 40

canonical uses of the subjunctive and works as a default case). Related to the sentence type is whether the words *que* “that.COMPL” or *si* “if” introduce the subjunctive verb (coded as *QUE* and, *SI*). The reason for including these two variables is that they are two of the most common triggers for the subjunctive, but have quite different functions, which means it is conceivable that they correlate with one or the other form. Two further contextual variables I included were *CATEGORY OF THE NEXT WORD* and *CATEGORY OF THE PRECEDING WORD*, these were extracted from the first letter in the POS tags provided by FreeLing. Additionally an *X* category was used for cases where there was no word or punctuation mark after or before the word. Finally I included the variable *LENGTH OF SENTENCE* (in number of words).

5 Productivity

Several authors have observed (Rojo 2008; Rojo and Rozas Vázquez 2014; Schwenter 2013; Wilson 1983) that the form *-ra* has been displacing the form *-se* during the last centuries. There even seem to be some contexts where *-ra* is acceptable and *-se* is not. An example of this is dialects where the imperfect subjunctive can replace the conditional (4), or in some journalistic styles:

- (4) a. Si yo fuera dueño de esta casa, yo estaría
 if I be.1SG.IMP.SBJ owner of this house I be.1SG.COND.PRES
 furioso
 furious
 “If I were the owner of this house, I would be furious”
 25
- b. Si yo fuera dueño de esta casa, yo estuviera
 if I be.1SG.IMP.SBJ owner of this house I be.1SG.IMP.SBJ
 furioso
 furious
 30
- c. *Si yo fuera dueño de esta casa, yo estuviese
 if I be.1SG.IMP.SBJ owner of this house I be.1SG.IMP.SBJ
 furioso
 furious
 35

This observation about the relative productivity of *-se* and *-ra* is confirmed by productivity metrics. Here I follow Gaeta’s (2007) approach (based on Baayen (1992)) for calculating the productivity of inflectional affixes. Gaeta proposes a method for comparing the degree of productivity of different affixes, while

controlling for the frequency of these. In Baayen’s original proposal (Baayen 1992) a *P*-index was calculated with the formula: $P(N) = h/N$, where *h* is the number of hapax legomena (single occurrences in a corpus) for a given affix, and *N* is the total number of tokens that appear with that affix. A larger *P*-index means a higher degree of productivity. In Gaeta’s approach the same *P*-index metric is used, but there is an additional sampling to control for frequency. The problem with the traditional *P*-index is that when comparing a high frequency affix with a low frequency one, the high frequency affix can be ranked as less productive, even in cases when the theory says this cannot be the case (see Gaeta (2007) for an example of this). Here I compare the totality of the -se examples, with the results from two sub-corpora for -ra. For the first sub-corpus I took the first 100 files of the CORLEC corpus (in alphabetical order), and for the second one I selected 100 random files. This way we can compare the productivity of both forms while controlling for frequency bias. The results can be seen in Table 1.

Table 1: Productivity ratings for -ra and -se.

Form	Sample	H	P-index	N
Ra	First 100 files	56	0.29	190
Ra	Random sample	54	0.29	189
Se	Whole corpus	46	0.23	191

As we can see, -ra is more productive than -se for both chosen sub corpora, which offers some quantitative confirmation for previous research on the topic that had also found -ra to be the more productive form. The fact that both samples of the corpus for -ra agree, means that the productivity of this form should be consistent throughout. Future research could look at the development of the *P*-index for both forms in a historical context.

6 Distribution of the alternation

After removing cases with incomplete information and some clear errors in the extraction, the total number of observations was 1,269, with some sentences containing more than a single occurrence. In agreement with DeMello (1993) and contradicting Gili Gaya (1983) the form -se (191 occurrences) is considerably less frequent than the form -ra (1078 occurrences). Other relevant proportions are presented in Table 2.

Table 2: Total number of occurrences, sentences and verbs with the forms *-se* and *-ra*.

	Se	Ra	total
Total cases	191	1,078	1,269
Number of sentences	171	911	1,081
Number of verbs	97	228	325

Figure 1 shows the proportions in which the alternation occurs with the variables TYPE, MODAL, NUMBER, QUE, PERSON and SI. In this figure we can see that both forms are almost identical except for the variables MODAL and TYPE.⁸ The morpheme *-ra* seems to appear with modals and verbs ending in *-er* more often than the morpheme *-se*. It is however likely that both of these variables are correlated to some degree because all modal verbs chosen end in *-er*.

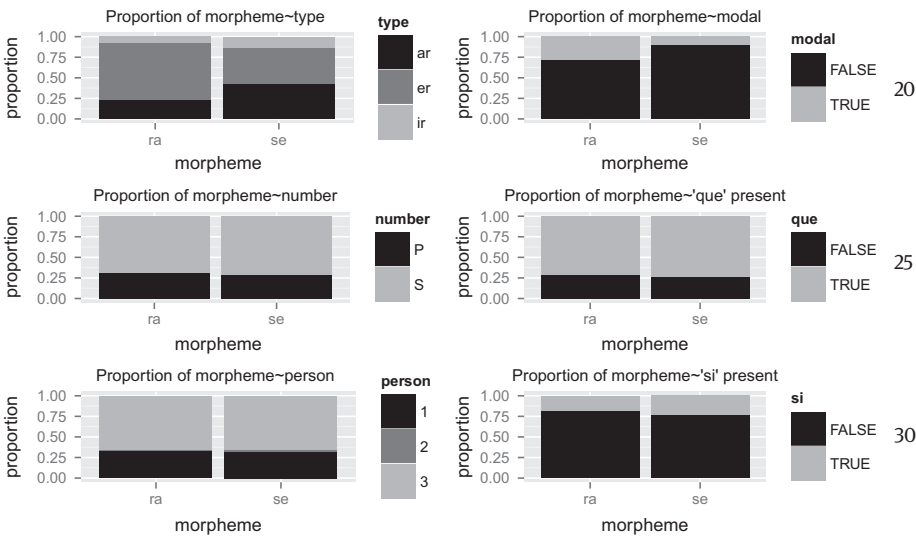


Figure 1: Proportions of the variables TYPE, NUMBER, PERSON, MODAL, QUE, SI for *-se* and *-ra*.

⁸ Statistical tests will be omitted in this section because the models presented in the next section are a better way of assessing the importance of each of these variables and their correlation with the forms of the alternation.

Figure 2 shows the proportions of realization of the subject and object. We can see that the differences in subject phrases are smaller than the differences for objects, but it is apparent that *-ra* appears with more sentences without overt subjects than *-se*. For objects the differences are larger. Most salient in these plots is that the form *-se* prefers noun phrases, while *-ra* shows almost the same preference for noun phrases and verb phrases.

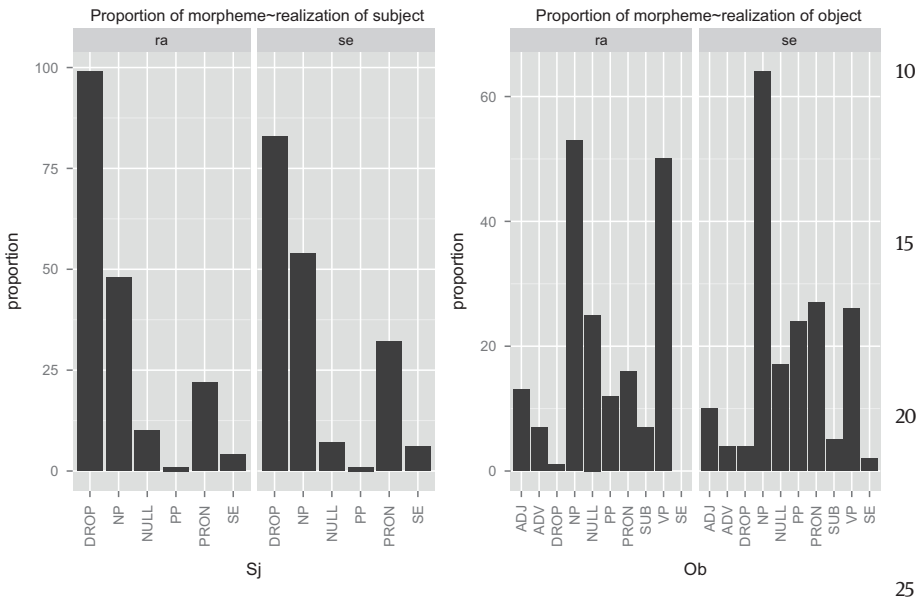


Figure 2: Proportions of the variables `REALIZATION OF SUBJECT` and `REALIZATION OF OBJECT` for *-se* and *-ra*. DROP=no overt subject near the verb, NP=noun phrase (with or without determiner), NULL=impersonal uses like existential *haber*, PRON=single pronoun (also relatives, demonstratives and numerals), SE=impersonal sentences with *se*, ADJ=bare adjectives and adjective phrases, ADV adverbial phrases, PP=prepositional phrases, SUB=subordinate clauses headed by a complementizer, VP=verb phrases without complementizer.

Figure 3 gives the proportions for animacy and definiteness of both subject and object (again, object here means any post verbal complement of the verb). As can be seen little difference in the animacy of subject and object, but there are noticeable differences in the definiteness of subject and object. The largest differences are between abstract (i. e., non NPs or PPs), definite and indefinite objects, but some difference between definite and indefinite subjects can also be observed.

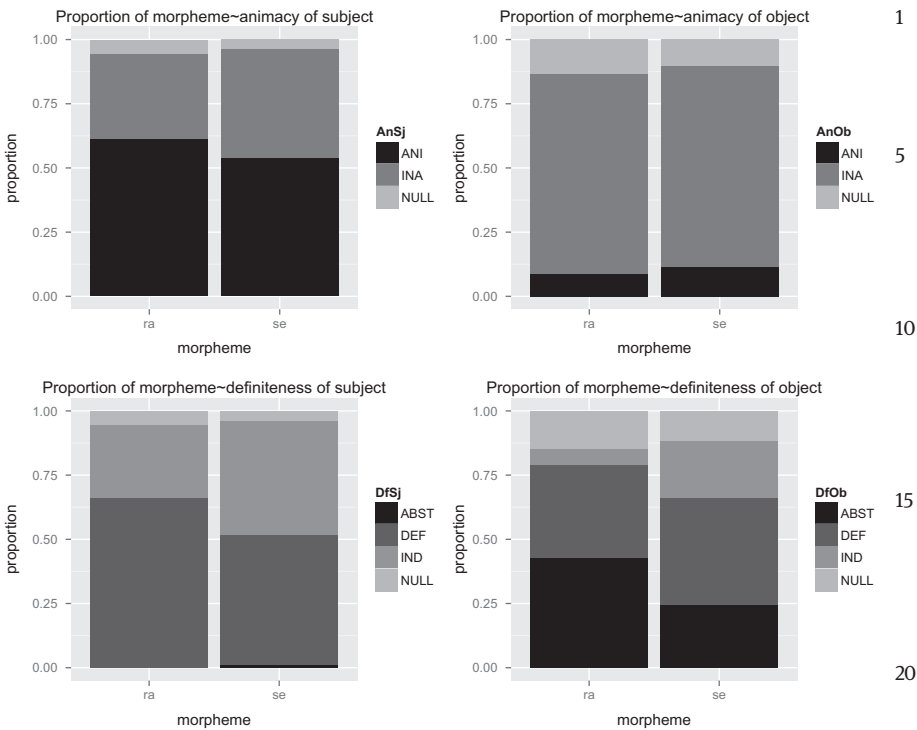


Figure 3: Proportions of the variables ANIMACY OF SUBJECT, ANIMACY OF OBJECT, DEFINITENESS OF SUBJECT, DEFINITENESS OF OBJECT for *-se* and *-ra*. NULL = no subject or object, ABST = for phrases different from NPs that work as the subject or first complement of the verb. DEF and IND are definite and indefinite subject and objects, both for NPs and NPs introduced by prepositions.

Figure 4 presents the types of sentences in which *-se* and *-ra* appear. We can see that the nonstandard uses of the subjunctive (adversative, indicative and temporal) are very uncommon with the imperfect subjunctive. Interestingly, there seems to be a reversal in proportion between potential and final sentences.

Figures 5 and 6 show the distributions of the grammatical categories of the preceding and following words.⁹ From both figures we can see that there does not seem to be much difference in the preceding grammatical category between both morphemes. The following grammatical category does show some differences,

⁹ The basic POS tags shown here can be found in the appendix. For a full list of what each POS tag means see <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html>.

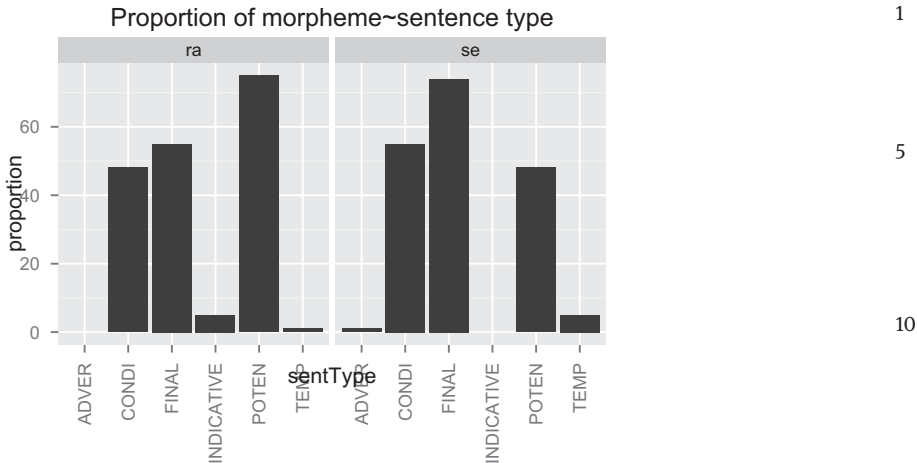


Figure 4: Proportions of types of sentences for *-se* and *-ra*. ADVER = adversative sentences, CONDI = conditional sentences. FINAL = subordinate sentences with *para* (etc.) that express intention, desired outcome or objective; INDICATIVE = any indicative use of the subjunctive, POTEN = default level and canonical use of the subjunctive, and TEMP = temporal uses of the subjunctive.

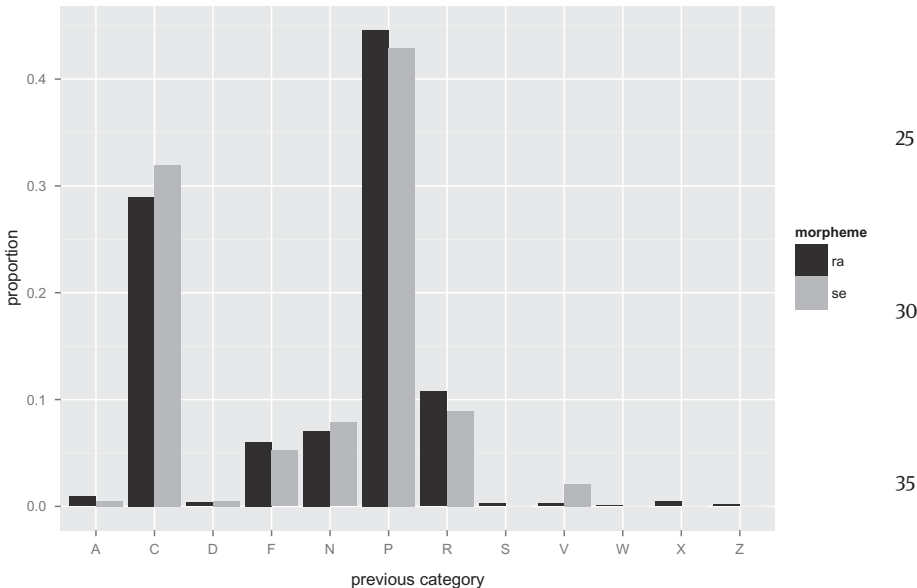


Figure 5: Proportion of preceding grammatical category present for *-se* and *-ra*.

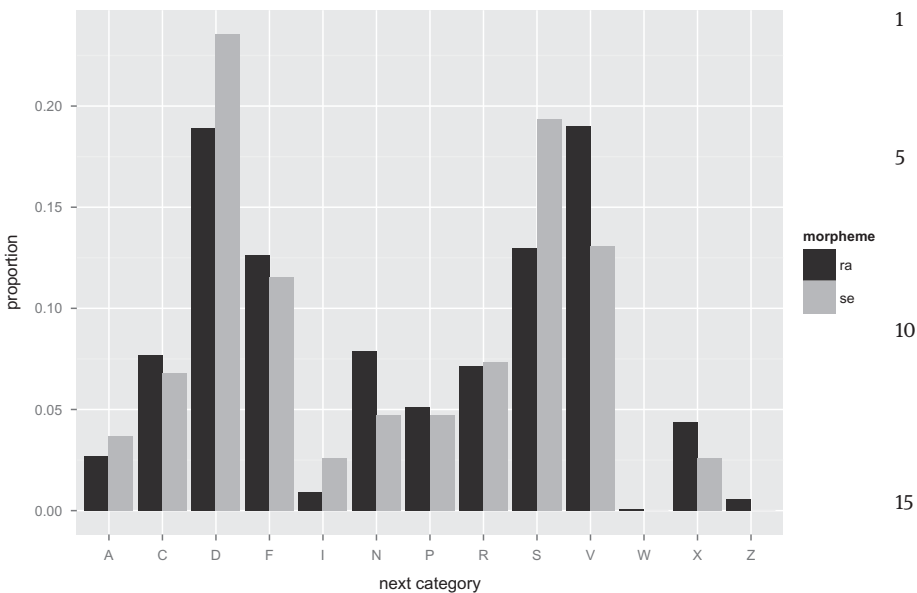


Figure 6: Proportion of next grammatical category present for *-se* and *-ra*.

mainly in prepositions (S), nouns (N), determiners (D) and main verbs (V), but the effect is not large enough to draw any conclusions yet. We will come back to the effects of preceding and next grammatical category in the following section.

The next factor considered was LENGTH OF SENTENCE. As mentioned above, there are repeated sentences in the data for the cases where a single sentence contains more than one case of the construction. Figure 7 shows the distribution of the length of sentence for each form considering only the manually coded cases.

We see that *-ra* seems to occur in slightly longer text units, but we also see that some of these text units are too long (over 100 words) to be actual sentences. It is possible that this difference is due to a difference in style (an effect by proxy), but since it is highly unlikely that this factor could have a direct effect, I will not consider it for the model.¹⁰

Finally, we can have a look at the variable VERB (considering all hits). If we examine the proportions of verbs for each form we can find that, not very surprisingly, *-ra* appears with considerably more verb types than *-se*, but *-se* also appears with some verb types that do not appear with *-ra*. The individual lists of verbs that exclusively appear with either *-se* or *-ra* are shown in Tables 3 and 4 respectively.

¹⁰ I am grateful to an anonymous reviewer for this observation.

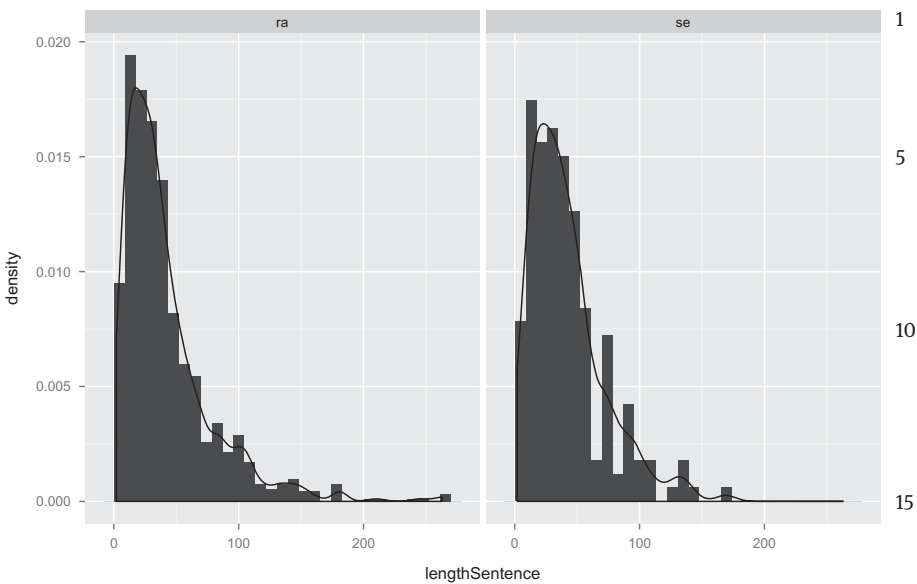


Figure 7: Histogram of length of sentence for -se and -ra.

Table 3: Verbs that appear with -se but not with -ra.

Verb	Gloss	Frequency	Proportion
aclarar	Clarify	2	0.0104712
desear	Wish	2	0.0104712
equivocar	Mistake	2	0.0104712
marcar	Mark	2	0.0104712
actuar	Act	1	0.0052356
adjudicar	adjudicate	1	0.0052356
alcanzar	Reach	1	0.0052356
alejar	move away	1	0.0052356
antojar	Fancy	1	0.0052356
aplicar	Apply	1	0.0052356
aprender	Learn	1	0.0052356
aprovechar	take advantage of	1	0.0052356
arrancar	pull out	1	0.0052356
asumir	Assume	1	0.0052356
ayudar	Help	1	0.0052356
calificar	Clarify	1	0.0052356
cifrar	Encode	1	0.0052356
compartir	Share	1	0.0052356

(continued)

Table 3: (continued)

1

Verb	Gloss	Frequency	Proportion
comprobar	Verify	1	0.0052356
concertar	agree on	1	0.0052356
concretar	make concrete	1	0.0052356
considerar	Consider	1	0.0052356
creer	Believe	1	0.0052356
derrumbar	Crumble	1	0.0052356
dirigir	Direct	1	0.0052356
encargar	order, ask	1	0.0052356
enfrentar	Confront	1	0.0052356
fallar	Fail	1	0.0052356
fijar	Fix	1	0.0052356
informar	Inform	1	0.0052356
jamar	Eat	1	0.0052356
lanzar	Throw	1	0.0052356
merecer	Deserve	1	0.0052356
moderar	Moderate	1	0.0052356
molestar	tease, bother	1	0.0052356
penetrar	Penetrate	1	0.0052356
precisar	make precise	1	0.0052356
profundizar	go in depth	1	0.0052356
reabrir	Reopen	1	0.0052356
realizar	Make	1	0.0052356
relajar	Relax	1	0.0052356
resolver	Resole	1	0.0052356
retomar	Retake	1	0.0052356
sentir	Sense	1	0.0052356
suministrar	Provide	1	0.0052356
valorar	Value	1	0.0052356

5

10

15

20

25

Table 4: Most frequent verbs that appear with *-ra* but not with *-se*.

30

Verb	Gloss	Frequency	Proportion
deber	Must	23	0.02133581
conocer	Know	7	0.00649351
ocurrir	happen	6	0.00556586
quitar	take away	6	0.00556586
acudir	go to	5	0.00463822
cambiar	change	5	0.00463822
contestar	answer	5	0.00463822
fallecer	die	4	0.00371058

(continued)

35

40

Table 4: (continued)

Verb	Gloss	Frequency	Proportion	
seguir	follow	4	0.00371058	
aparecer	appear	3	0.00278293	5
coger	take, grab	3	0.00278293	
comprar	buy	3	0.00278293	
dedicar	dedicate	3	0.00278293	
desaparecer	disappear	3	0.00278293	
explicar	explain	3	0.00278293	
funcionar	function	3	0.00278293	10
jugar	play	3	0.00278293	
mover	move	3	0.00278293	
pedir	ask for	3	0.00278293	
preguntar	ask	3	0.00278293	
presentar	present	3	0.00278293	
reconocer	recognize	3	0.00278293	15
usar	use	3	0.00278293	
vender	sell	3	0.00278293	
abrir	open	2	0.00185529	
acercar	move closer	2	0.00185529	
arreglar	repair	2	0.00185529	
atender	help	2	0.00185529	20
caber	fit	2	0.00185529	
caer	fall	2	0.00185529	
comentar	comment	2	0.00185529	
comenzar	begin	2	0.00185529	
constituir	constitute	2	0.00185529	
cuidar	take care of	2	0.00185529	25
esperar	wait	2	0.00185529	
establecer	establish	2	0.00185529	
estudiar	study	2	0.00185529	
existir	exist	2	0.00185529	
financiar	finance	2	0.00185529	
leer	read	2	0.00185529	30
mandar	order, send	2	0.00185529	
morir	die	2	0.00185529	
nacer	be born	2	0.00185529	
notar	notice	2	0.00185529	
ofrecer	offer	2	0.00185529	
olvidar	forget	2	0.00185529	35

A detailed analysis of verb collexemes of the imperfect subjunctive constructions will be presented in Section 8, but both these tables already suggest that there are lexical preferences associated with either form. We can see that *deber* (“must”), a modal verb, never appears with *-se*, perhaps indicating that modality of the verb might play a role in the selection of one form or the other. This is consistent with the proportions of modals we saw before, but the results must be tested for significance.

Just looking at raw frequencies is not enough to determine whether there are significant correlations between these variables and the *-se/-ra* alternation. Statistical testing of each individual variable would also be of little help because such a procedure cannot take into account interactions between the variables, and multiple testing reduces the reliability of each individual test. To address this problem we now turn to multifactorial methods.

15

7 Multifactorial interactions

The use of multifactorial methods and machine learning algorithms for predicting alternations is a relatively recent development in corpus linguistics that started with studies by Gries (2003) and Bresnan et al. (2007), and these methods are becoming increasingly popular in the field of Cognitive Linguistics and Corpus Linguistics (Janda 2013). In most approaches, researchers try to find the best fit by the backwards elimination of factors based on *p*-values. Here I take a slightly different approach. The main reason is that the algorithm that I will be using, Naive Discriminative Learning (NDL) does not allow for backward elimination of factors based on *p*-values, instead I will focus mostly on the C score and on AIC scores of the models for model selection (see Johansson (2011) for an argument against the use of *p*-values for model selection).

30

7.1 Initial considerations

The first issue to be considered regarding regression models is which factors should be included in the initial model. The natural choice are the factors already discussed in the previous section: VERB, PERSON, NUMBER, LENGTH OF SENTENCE, MODAL, SI, QUE, PRECEDING CATEGORY, NEXT CATEGORY, animacy of the subject (ANSJ) and object (ANOB), definiteness of the subject (DFSJ) and object (DFOB), the realization of subject (SJ) and object (OB), and the SENTENCE TYPE.

40

The second issue that requires consideration is which kind of model should be fitted to the data. The most widely used machine learning algorithm for the purpose of linguistic data analysis is logistic regression (with and without random effects). Other popular methods include partition trees and random forest. Finally, a new model that has shown very promising results is Naive Discriminative Learning (Baayen 2010, 2011; Baayen et al. 2011a, 2011b, 2013). The main advantage of the latter model is that it is not based on abstract equations (like logistic regression) or a black box (like Random Forest), but on work on classical conditioning and discriminative learning (Rescorla and Wagner 1972), which has proven to be an excellent model for animal and human learning (Miller et al. 1995).¹¹ In what follows I will use Naive Discriminative Learning for most of the models.

Naive Discriminative Learning is based on the Rescorla–Wagner equations. The basic idea behind this model is that animals learn in a cue–outcome fashion. If a cue is present when an outcome is seen, then the value of that cue (the association between the outcome and the cue) increases; when a cue is absent when an outcome is seen, then the value of that cue decreases. The Rescorla–Wagner equations describe how the association between outcome and cues changes by each observation. The equations that describe the model are as follows:

$$\Delta V_x^{n+1} = \alpha_x \beta (\lambda - V_{tot})$$

$$V_{tot} = V_x^n + \Delta V_x^{n+1}$$

where ΔV_x^{n+1} is the change in association of X. α and β are fixed parameters bounded between 0 and 1, usually set at 0.1. λ is a fixed value denoting the maximum association strength for the unconditioned stimulus, usually set at 1. V_{tot} is the total sum of all association strengths, and V_x is the current association strength (for a more detailed explanation of how the model works see Baayen (2011)).

The result of the model is a set of weights for each cue for each outcome. Weights can be positive or negative (depending on whether a cue positively or negatively correlates with an outcome), and are normally bounded between 1 and –1. Cues with larger weights (relative to other cues) for a given outcome will be said to be strong(er) predictors of that outcome.

¹¹ Notice that although NDL offers the advantage of providing a clear way in which speakers could be making the generalizations we find in the data, I do not mean to imply that all generalizations found have psychological reality. Some of the generalizations will simply be patterns in the corpus. Psycholinguistic work would be required to determine which of those patterns have psychological reality.

For model assessment I will mainly use the Area under the Roc Curve value (C). 1
The C score can range from 0 to 1, with 1 being a perfect model fit, and 0 a perfectly
wrong model fit. Models with values from 0.5 to 0.6 are considered to be bad
models (they perform no better than chance), those from 0.6 to 0.75 are considered
to be decent models, those from 0.75 to 0.9 are considered to be good models, those 5
from 0.9 to 0.97 are considered to be very good models, and those from 0.97 to 1.0
are considered to be excellent models.

7.2 Morpho-syntactic and discourse factors

10

The smallest model that best fits the data has the following formula: morpheme
~ modal + DfSj + DfOb + sentenceType + verb (Model A). This formula means
that we are taking morpheme as the dependent variable, and the variables after
the ~ as predictors. Other predictors, especially number and person did not 15
appear to be relevant for the model (i. e. they had no appreciable effect on the C
score of the model). The confusion matrix for this model is shown in Table 5.
We can see that the model fits the data very well, with very few errors.

Table 5: Confusion matrix for model A.

20

Confusion Matrix		
Reference	Prediction	
	Ra	se
Ra	160	24
Se	33	150
Accuracy: 0.8446		
C score: 0.9166		

25

By far the strongest predictor was VERB, which suggests very strong lexical
preferences in the construction. Since Section 8 will deal exclusively with the
issue of lexical effects, I will not go into a detailed discussion of this predictor
here (the next section also offers some additional details on how strong VERB
actually is and why). 30

The best other individual predictors for -ra and -se are given in Figures 8
and 9. From these figures we can see that these predictors are relatively weak.
The strongest of those predictors for both -se and -ra was DEFINITENESS OF THE
SUBJECT, with null subjects predictive of -ra and abstract subjects (those different
from NPs) predictive of -se. We see as next best predictor of both the sentence 40

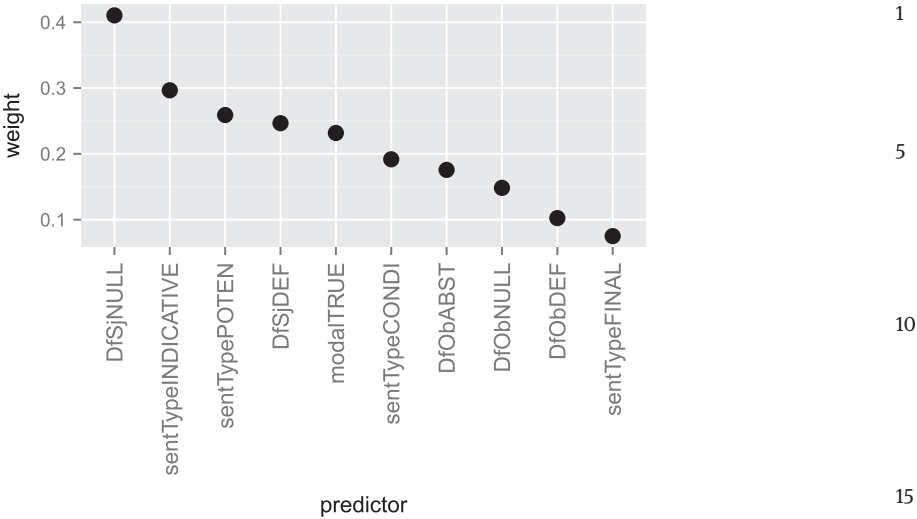


Figure 8: Best 10 predictors for *-ra* excluding *VERB*. In order from left to right: definite subjects, indicative sentences, potential sentences, definite subjects, modal verb, conditional sentences, abstract objects, null objects, definite objects, final sentences.

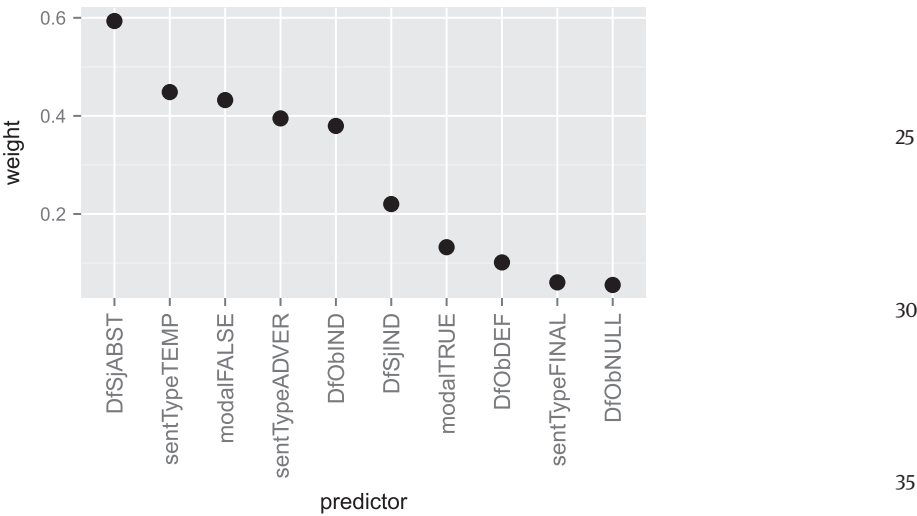


Figure 9: Best 10 predictors for *-se* excluding *VERB*. In order from left to right: abstract subjects, temporal sentences, no modal verb, adversative sentences, indefinite objects, indefinite subjects, modal verbs, definite objects, final sentences, null objects.

type, with adversatives and temporal sentences predictive of *-se* and indicative 1
and potential sentences predictive of *-ra*. Also interesting is MODAL, which seems
to be a moderately strong predictor for *-ra*. This is consistent with the previously
observed differences in the use of modals between both forms. Regarding
SENTENCE TYPE, we see that most of the levels selected are those that appear only 5
a few times (temporal, adversative and indicative), which would be expected
and is not very informative. The exception is potential sentences, which are
preferred by *-ra*. Since this level represents the default or “normal” use of the
subjunctive (outside the conditional), it seems that there is a degree of specia-
lization of *-se*. A much larger corpus would be needed to assess the importance of 10
the other levels.

A detailed interpretation for each single level of each predictor is not easy (and
because of their low scores not very enlightening), but from these results it is clear
that the strongest predictors of both *-se* and *-ra* are not grammatical levels of the
verb-related variables, but lexical ones (related to the verb itself), or levels of 15
context-related variables. This contradicts the results by Schwenter (2013).¹²

7.3 Model evaluation and overfitting

Although the previous model achieved high accuracy, it is important to evaluate 20
how much the patterns observed are specific to this particular data-set, and
which effects are more likely to be part of the alternation as a whole. There are
several techniques to test this. The first one I will employ is bootstrapping the
model by splitting the data into training and testing portions, and repeating the 25
process multiple times (30 in this case), and the second one is using machine
learning algorithms that are less prone to overfitting.¹³

7.3.1 Cross-validation and model selection

The results from bootstrapping Model A are presented in Table 6. We can see
in Table 6 that there is a significant drop in accuracy and C score, but never-
theless the model seems to retain some predictive capability above random 35

¹² This does not mean, however, that his results are wrong. The difference could be due to the
use of different corpora or to the way he collected the data. A direct replication attempt would
be necessary to evaluate Schwenter's results.

¹³ Overfitting means that a model fits a particular data-set very well, but it does not work as
well on new data. 40

Table 6: Mean confusion matrix for bootstrap of model A.

1

Confusion Matrix			Prediction
Reference	Ra	se	
Ra	4.97	1.57	
Se	3.00	3.47	
Mean Accuracy: 0.6487 %			
Mean C score: 0.7173			

5

10

chance. This suggests that the model might be capturing some real correlations in the data, although it is not as powerful as initially thought.

We can follow up and ask where the overfitting is coming from. The most likely candidate for an explanation is *VERB*. This variable is a good predictor 15 because it has many levels, which means it has more chances of establishing a correlation with the dependent variable (Kapatsinski 2013).¹⁴ A way of improving this result is by reducing the number of levels of this variable. We can achieve this by assigning the same value (“default” in this case) to all levels that appear in the data set less frequently than an arbitrary threshold (here set to 3). The 20 resulting variable is a reduced predictor with only 13 levels (including “default”). The resulting model (Model A2) has and even worse performance than Model A, but still manages to discriminate correctly many cases as can be seen in Table 7.

25

Table 7: Confusion matrix for model A2.

Confusion Matrix			Prediction
Reference	Ra	se	
ra	132	52	
se	56	127	
Accuracy: 0.7057			
C score: 0.7734765			

30

35

We can conclude that much of the overfitting was due to the multiple levels of *VERB*, but even after controlling for this, the model still managed to correctly classify a good

40

14 I am grateful to one of the anonymous reviewers for pointing this out.

number of observations. However, these results do raise the question of whether all the predictors tested are “significantly” better than chance.

NDL has no method for calculating p -values, which means we cannot directly test this in the traditional way. This is not too much of a problem, however, as the use of p -values for model selection has been called into question (see for example Johansson (2011)). An alternative is using the likelihood ratio for AIC scores (Glover and Dixon 2004; Johansson 2011).

For this technique we calculate the bits of evidence for a given restricted model contained within another model, that is, a model that has fewer predictors than another one. For this we use the equation:

$$(AIC(R1) - AIC(R2)) * \log_2(\exp(1))$$

where $R1$ is the restricted model and $R2$ is the unrestricted model. The sign of the result indicates whether the data provides evidence for the restricted model (–) or for the unrestricted model (+), while the magnitude is the strength of the evidence. We can use this method for model selection. We first calculate the evidence for or against all predictors, we then remove the predictor against which we have the strongest evidence (if there is one), and then repeat until we have a model completely favored by the data.

Table 8 presents the resulting bits of evidence for the predictors. We can see that this process eliminates `VERB` and `SENTENCE TYPE` as predictors justified by our data, and that the evidence for `DEFINITENESS OF SUBJECT` is weak. Meanwhile, both `MODAL` and `DEFINITENESS OF OBJECT` are justified by our data according to this method. The cases of `VERB` and `SENTENCE TYPE` are interesting. They seem to be overall bad predictors, but have a couple of levels that are strong predictors. This is particularly prominent in the case of `VERB`. It is a bad predictor in the sense that many levels (many verbs) are not strongly associated with any of the two forms, but it is a good predictor in the sense that some verbs are in fact very strongly associated to one of the two forms, as we will see in the collostructional analysis.

Table 8: Individual predictor performance according to bits of evidence.

Factor	Iteration	Bits of evidence
Verb	1	–56.17
Sentence Type	2	–22.63
Definiteness of Subject	3	0.15
Definiteness of Object	3	12.25
Modal	3	7.38

7.3.2 Random forest1

The second technique we can use to evaluate the model is to use Random Forest (Breiman 2001; Liaw and Wiener 2002), which is a lot less prone to overfitting than other classification algorithms because it splits the data during training. 5

There are two main Random Forest algorithms: the original one proposed by Breiman (2001) and implemented in the random Forest package, and an improved version which is less prone to overestimating the effect of predictors with many levels (Hothorn et al. 2006; Zeileis et al. 2008), implemented in the party package. Because of the concerns over some variables having too many 10 levels, I fitted the model with the cforest_unbiased option to control for the difference in predictors. To test the previous claim that number and person had no impact on the model, I also added these two predictors to the random forest. The results can be seen in Table 9.

15

Table 9: Confusion matrix for the random forest model.

Confusion Matrix		
Reference	Prediction	
	Ra	se
ra	125	37
se	59	146
Accuracy: 0.737 %		
C score: 0.741		

20

25

Overall, the random forest model produces very similar results (in the sense of classification accuracy) to the modified NDL model (Table 7), or the cross-validated NDL model (Table 6). Another advantage of random forest is that we can test for 30 predictor importance (using the conditional option because MODAL and VERB are highly correlated). We see in Table 10 that the model found definiteness of the object, sentence type and definiteness of the subject to be better predictors than the verb. This is most likely, as we saw before, because VERB is only a very good predictor in a few specific cases of verbs that clearly prefer one form or the other. We can also see some 35 confirmation that person and number are the weakest predictors.

We can conclude that although our model is not a perfect fit, there are contextual factors (definiteness of the subject and object, and the sentence type), as well as lexical effects of the verb, that are weakly correlated with the 40 forms in the *-se/-ra* alternation.

Table 10: Best predictors for the random forest model.

Predictor	Mean Decrease Accuracy
DfOb	0.0355
sentType	0.0104
DfSj	0.0101
modal	0.0054
verb	0.0035
number	0.0017
person	−0.001

8 Collostructional analysis

Finally, to investigate in depth the lexical preferences of each morpheme I conducted a collostructional analysis (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004). The idea behind collostructional analysis is that just as it is possible to measure the strength of attraction between a word and its collocates within a defined span, it is also possible to measure the attraction between a construction and the lexemes that occur in a fixed structural position of that construction. For this analysis I focused only on the position of the verb (X in the schema presented in (2)) and not on positions in the sentence. I use the complete data-set as specified in Section 4 for this part of the analysis.

8.1 Attracted collexemes

First we look at the 20 collexemes that are most strongly attracted to both *-ra* (Table 11) and *-se* (Table 12). The first interesting fact that can be observed is that the top three positions for *-ra* are occupied by verbs that can typically be used as modals: *querer* “want”, *poder* “can” and *deber* “must”.¹⁵ In contrast, for *-se* we find that the construction does not attract any of these modal verbs. We can see that the difference in collexemes is quite strong, there is no overlap in these first 20 verbs. Another important point is the strength of attraction. If we compare the strength of attraction of the first three collexemes for *-ra* we can see that it is considerably stronger than all other collexemes for *-ra*, suggesting that these are

¹⁵ To be absolutely sure that all cases of *querer* are in fact modal uses, a manual coding of the whole corpus would be necessary. While this is not feasible due to the size of the corpus, in a random sample of ten sentences containing the verb, only one was not a modal use of it.

Table 11: First 20 attracted collexemes for *-ra*. 1

N	Verb	Gloss	Co-occurrences	Expected Frequency	Verb Frequency	Fisher's <i>p</i>	
1	querer	want	114	12.164734	2375	1.954e-65	5
2	poder	can	103	25.455981	4781	2.721e-29	
3	deber	must	23	4.263007	747	3.091e-10	
4	acudir	come to	5	0.323384	59	2.900e-05	
5	fallecer	die, perish	4	0.167837	32	4.001e-05	
6	Ser	be	184	131.988916	26637	7.163e-05	
7	contestar	answer	5	0.898289	155	2.523e-03	10
8	financiar	pay for	2	0.066059	13	2.668e-03	
9	quitar	take away	6	1.567549	268	5.826e-03	
10	orear	air	1	0.000000	1	5.981e-03	
11	transfundir	transfuse	1	0.000000	1	5.981e-03	
12	pinchar	poke	2	0.108096	20	6.320e-03	
13	Usar	use	3	0.419983	73	9.715e-03	15
14	quedar	remain	13	5.860728	999	1.037e-02	
15	aguar	ruin	1	0.006011	2	1.193e-02	
16	apalear	beat	1	0.006011	2	1.193e-02	
17	desbancar	unseat	1	0.006011	2	1.193e-02	
18	desplomar	fall	1	0.006011	2	1.193e-02	
19	fusilar	execute, shoot	1	0.006011	2	1.193e-02	20
20	constituir	constitute	2	0.174155	31	1.481e-02	

the most central to the meaning of the construction. Also, if we examine more closely the collexemes for *-se* we can see that the strength of attraction is quite weak, and the actual number of co-occurrences of these top 20 collexemes is not greater than three. This suggests that these numbers are more likely due to chance than any actual semantic effect, but because of the sparsity of the data we cannot be sure. We can only be confident that *-ra* strongly attracts modal verbs while *-se* does not show any clear preferences. 30

8.2 Repelled collexemes

We can also take a look at the repelled collexemes, that is, the lexemes that we find with a frequency lower than expected for *-ra* (Table 13) and *-se* (Table 14). The first interesting observation is that the verb *ir* “go” (also as future tense auxiliary: *voy a dormir* “I am going to sleep”) is strongly repelled by *-ra* and it also appears on top (although with a weak effect) for *-se*. The most likely explanation is that the whole abstract construction in (2) is disliked with the periphrastic future form with *ir*. 35
40

Table 12: First 18 attracted collexemes for *-se*.

1

N	Verb	Gloss	Co-occurrences	Expected Frequency	Verb Frequency	Fisher's <i>p</i>
1	disparar	Shoot	2	0.032700	33	0.0005826
2	reabrir	Reopen	1	0.000000	1	0.0010649
3	aclarar	Clarify	2	0.053797	53	0.0014990
4	antojar	Fancy	1	0.001060	2	0.0021286
5	cifrar	Encode	1	0.002121	3	0.0031912
6	desear	Desire	2	0.088607	86	0.0038838
7	equivocar	Mistake	2	0.094936	92	0.0044292
8	escribir	Write	3	0.321073	309	0.0045369
9	derrumbar	Crumble	1	0.004242	5	0.0053130
10	marcar	Mark	2	0.106539	103	0.0055151
11	suministrar	Provide	1	0.005302	6	0.0063721
12	concertar	agree on	1	0.006363	7	0.0074302
13	levantar	Lift	2	0.129745	125	0.0080112
14	adjudicar	adjudicate	1	0.007423	8	0.0084871
15	retomar	Retake	1	0.010604	11	0.0116511
16	estallar	Burst	1	0.012725	13	0.0137547
17	Profundizar	go in depth	1	0.012725	13	0.0137547
18	concretar	fix, set	1	0.015906	16	0.0169018

20

If we examine the eight cases of *ir* that occur with this construction, only three are clearly cases of *ir a* as a future marker, which suggests that in fact the periphrastic future form is repelled by the construction. It is also interesting that *haber*, which is also used for periphrastic tenses (perfect and pluperfect), is repelled by *-ra*. What both these repelled collexemes suggest is that the whole construction repels periphrastic verb conjugations.

25

Another apparent pattern we find for *-ra* is that quite a few of the repelled verbs are expression verbs or psychological verbs: *saber*, *pensar*, *creer*, *decir*, *hablar*, *entender*, *considerar*. One possible explanation for these anti-collocations is that the construction simply repels the semantic field of expression and know/think verbs. It is however not clear at all why this should be the case. A different possibility is that verbs like *decir*, *creer* and *pensar* are often associated with subjectivity or evidentiality in the sentence, and the actual effect is not so much by the semantic field of these verbs but by the modality usually expressed by these kind of verbs. With the current data it is not possible to distinguish between these two explanations.

30

35

For *-se* there are no repelled lexemes that reach significance ($p < 0.05$). In these cases *ir* is probably related to the same issues discussed for *-ra*, but an interpretation for *hacer* is less clear. The fact that all *p*-values are too large, and

40

Table 13: First 20 repelled collexemes for *-ra*.

1

N	Verb	Gloss	Co-occurrences	Expected Frequency	Verb Frequency	Fisher's <i>p</i>	
1	Ir	go	5	45.435	7592	1.553e-13	5
2	haber	have (auxiliary)	65	91.692	16283	6.534e-03	
3	saber	know	9	19.808	3329	1.141e-02	
4	mirar	look	1	6.828	1137	1.802e-02	
5	valer	be worth	1	4.604	767	9.957e-02	
6	pensar	think	1	4.803	800	1.012e-01	
7	decir	say	35	46.022	7941	1.279e-01	10
8	creer	believe	0	2.340	389	1.827e-01	
9	Fijar	fix	0	2.581	429	1.926e-01	
10	hablar	talk	5	8.707	1459	2.999e-01	
11	Vivir	live	1	2.939	490	3.808e-01	
12	encontrar	find	1	3.144	524	3.854e-01	
13	entender	understand	1	2.807	468	5.372e-01	15
14	empezar	begin	2	3.483	582	5.936e-01	
15	Ver	see	21	24.175	4119	6.051e-01	
16	considerar	consider	0	1.029	171	6.310e-01	
17	realizar	make, do	0	1.077	179	6.318e-01	
18	llegar	arrive	4	5.946	996	6.753e-01	
19	intentar	try	1	2.038	340	7.283e-01	20
20	producir	produce	1	2.086	348	7.287e-01	

that the differences between observed co-occurrence and expected co-occurrence are too small means that it is quite possible that the distribution of most lexemes given in Table 14 is a product of chance alone. However, we find some interesting overlap with the lexemes repelled by *-ra*: *ir*, *saber*, *mirar*, *ver* and *decir*. This suggests that the construction as a whole, independently of whether it is instantiated as *-se* or *-ra*, has lexical dispreferences regarding these verbs. Even more interesting is that we also find some overlap with the collexemes attracted by *-ra* and repelled by *-se*, namely *querer*. This indicates not only very strong lexical preferences by both forms but distinctive lexical preferences.

8.3 Contrastive collexemes

35

We can also contrast the collexemes for *-se* and *-ra* by evaluating whether the proportion observed for each verb for each form is likely to be due to chance (that is, as it would be expected from the proportion of both constructions), or if there is likely to be a preference. This method simply tests the null hypothesis

Table 14: First 20 repelled collexemes for *-se*.

1

N	Verb	Gloss	Co-occurrences	Expected Frequency	Verb Frequency	Fisher's p
1	Ir	go	3	7.962829	7592	0.09562
2	Hacer	do, make	2	5.840652	5539	0.13556
3	Saber	know	1	3.529083	3329	0.27326
4	Decir	say	5	8.238338	7941	0.36601
5	Querer	want	1	2.517441	2375	0.52673
6	Dar	give	1	2.532287	2389	0.52687
7	Mirar	look	0	1.212045	1137	0.63769
8	Ver	see	3	4.318752	4119	0.80445
9	Ampliar	expand	0	0.026650	25	1.00000
10	Cocinar	cook	0	0.026650	25	1.00000
11	Comer	eat	0	0.321933	302	1.00000
12	comercializar	commercialize	0	0.008528	8	1.00000
13	Comprar	buy	0	0.495691	465	1.00000
14	Conocer	know	0	0.652394	612	1.00000
15	Depender	depend	0	0.178022	167	1.00000
16	Deriver	derive	0	0.027716	26	1.00000
17	Distribuir	distribute	0	0.027716	26	1.00000
18	Echar	throw out	0	0.272897	256	1.00000
19	Enchufar	plug in	0	0.013858	13	1.00000
20	Escoger	pick	0	0.027716	26	1.00000

that the distribution of each verb would be the same for both forms if there were no lexical preference. Using Fisher's exact test we can test the difference of each proportion and then rank them accordingly. The ten most distinct collexemes are shown in Table 15.

Table 15: Contrastive collexemes for *-se* and *-ra*.

30

N	Verb	Gloss	-ra	-se	Verb Frequency	Fisher's p
1	querer	Want	114	1	2375	0.0000006745
2	poder	Can	103	6	4781	0.0040894370
3	pensar	Think	1	3	800	0.0123960420
4	llegar	arrive	4	4	996	0.0223242134
5	aclarar	clarify	0	2	53	0.0229566898
6	desear	desire	0	2	86	0.0229566898
7	equivocar	mistake	0	2	92	0.0229566898
8	marcar	mark	0	2	103	0.0229566898
9	deber	must	23	0	747	0.0375682154
10	escribir	write	3	3	309	0.0487333244

40

This table supports what we had already observed from the collostructional analysis, namely that *querer*, *poder* and *deber* are strong indicators for *-ra*, but it also tells us that the other seven verbs are all tipped in favor of *-se*. We see that some of the verbs that we already saw in the top 20 collexemes for *-se* appear here, namely *aclarar*, *desear*, *equivocar*, *marcar* and *escribir*, and we also see *llegar*, which was in the list for repelled collexemes for *-ra*. Because *-se* is a lot less frequent than *-ra* we would not expect to see verbs like *llegar* or *escribir* occurring with the same raw frequency with *-se* and *-ra*, and we would definitely not expect to see verbs like *pensar* being more frequent with *-se* than with *-ra*. This converging evidence strongly indicates again that there are clear and distinctive lexical preferences that distinguish *-se* from *-ra*, even though it is not clear what the criteria are behind the collexemes attracted to *-se*.

9 Discussion

The basic frequency counts of both forms confirm previous claims that *-se* is less frequent than *-ra*, at least in spoken language (DeMello 1993; Rojo 2008). Similarly, the productivity measures show that *-ra* is also more productive than *-se*, which agrees with the previous claims that *-ra* is taking over' and replacing *-se*.

As we have seen, the Naive Discriminative Learning model shows that some discourse and context factors are weakly but significantly correlated with the *-se/-ra* alternation, while the core grammatical factors NUMBER and PERSON are either too weak, or not correlated at all (counter to Schwenter (2013)). These effects remain present after controlling for overfitting. The model also presents evidence for strong lexical effects, both in the lexical choices of individual verbs, and in the overall preference of *-ra* for modal verbs.

From the collostructional analysis we can conclude that the construction has strong lexical effects. The strongest effect we found was that the form *-ra* attracts modal verbs but *-se* does not, and even possibly repels them. We can also be confident that the general construction repels the verb *ir* ("go" as future marker), most likely because it repels constructions with the periphrastic future tense, and possible other periphrastic constructions with *haber* ("have"). Finally, we also saw that the two verbs that are most distinctive between both constructions are *querer* ("want") and *poder* ("can"). All these facts very strongly support the case for pragmatic difference between both forms, but also for some pragmatic similarities.

These results are directly relevant for the constructional analysis proposed for this alternation. Because the model only reached a moderate accuracy, and this

accuracy dropped significantly in cross-validation and with Random Forest, we can
conclude that there is in fact a very close relation between both forms, and speakers
do use them interchangeably to a large extent. More specifically, neither *NUMBER* nor
PERSON helped distinguish between both forms. This can be understood within the
proposed framework of construction grammar if we allow the activation of these
factors to occur at the level of the more general construction (2), while the activation
of the lexical items and discourse factors is closer to the activation of one of the
concrete schemas in (3). We can then propose an updated and more detailed
representation of these constructions in (5) and (6):

- (5) $[[X_{vi}] -Y_{se/ra} [\text{PERSON}] [\text{NUMBER}]]_v \leftrightarrow [\text{SEM}_i \text{ in imperfect tense subjunctive} + \text{PRAG}_1]$ 10

Based on the results of the models we can propose that the *NUMBER* and *PERSON*
constructions are instantiated in the abstract construction in (5). This means that
at the level of (5) both *NUMBER* and *PERSON* are free slots in the constructions. The
more specific constructions for *-se* and *-ra* would be the following:

- (6) a. $[A_{vi(j)} -ra_j + \text{PERSON/NUMBER}]_v \leftrightarrow [\text{SEM}_i \text{ in imperfect tense subjunctive} + \text{PRAG}_1 + \text{PRAG}_j]$ 20
b. $[B_{vi(k)} -se_k + \text{PERSON/NUMBER}]_v \leftrightarrow [\text{SEM}_i \text{ in imperfect tense subjunctive} + \text{PRAG}_1 + \text{PRAG}_k]$

where A and B stand for concrete lexical choices (in contrast to the free slot in (5))
that are partially linked to the specific form *-se* or *-ra* (this represents the lexical
preferences of *-se* and *-ra*), PRAG_k and PRAG_j are elements of discourse related
to the complements of the verb and possibly the sentence type where the sub-
junctive appeared. At this level both *PERSON* and *NUMBER* are not free slots, but
are inherited from the more abstract construction in (5) (and the individual
constructions for number and person). For PRAG_1 , discourse preferences common
to both forms, it has not been possible to find any direct associations. Nevertheless,
some features like the dispreference of some periphrastic constructions by both
forms, and the fact that conditional sentences are used equally for both forms,
can be seen as common elements of *-se* and *-ra*. Understanding how
definiteness of the object and subject play a role is less straightforward and
requires further research. It is possible that this variable is only acting as proxy
for some semantic effect.

Notice that the analysis presented in (5) is actually reminiscent of a mor-
phomic level as proposed by Aronoff (1994). In this analysis, there is a level that
lies directly in the morphology and is not accessible to the syntax. It might turn

out that these intermediate abstract levels are in fact required for enforcing (quantitative) locality effects in morphology (see also O'Neill (2014)). Independently of whether the construction morphology analysis presented in this paper is correct, the results of the models and the constructional analysis are empirical evidence that lend some support for a constructional approach to verbal inflection where grammatical constructions combine with lexical constructions to produce conjugated verbs. We need a constructional view because the schema that produces the imperfect subjunctive is not only associated with a specific grammatical meaning, but it also exhibits very complex distributional patterns that need to be represented and associated with it. The emergence of these patterns is best explained from a usage-based perspective where each exemplar counts, and each exemplar can be richly represented including the context it appeared in.

10 Final considerations

The main result of this study is that the *-se/-ra* alternation is not completely unpredictable from the morpho-syntactic and discourse context, and that the null hypothesis that both forms are in complete free variation is most likely wrong. However, it must be emphasized that the models used only show the existence of correlations between the predictors and the response variable, and that this does not imply causation. Since we do not have a good understanding of how speakers actually plan and produce sentences, how they choose what to say and how to say it, it is not possible to give a detailed account of exactly what these correlations mean, or how they actually work in production.

It must be noted that there is no native' implementation of stochastic processes in construction grammar. However, cognitive versions of construction grammar assume that domain general cognitive processes are responsible for, and interact with, constructions. This means that an NDL mechanism could be part of the whole system and operate at different levels of granularity and abstractness (here lies the advantage of NDL over many other machine learning algorithms).

An issue that is always present when modeling alternations in language is that it is not possible to know beforehand how much variability we should be able to account for with our models, and how much variability should not be possible to model, as it is likely that a degree of variation is just probability matching (Kapatsinski 2010, 2014). We do not know a priori how much freedom speakers actually have when they choose one form or the other, and how much is determined by context. This means that it is impossible in principle to ever

know if the statistical model we chose reached ceiling or if there are other still 1
unknown predictors that, if included, would increase model performance. All we
can say for certain is that the use of *-se* and *-ra* is not completely random, and
that there are at least some real correlations with the factors mentioned.

Old issues that appeared to have been settled with the use of traditional 5
linguistic methods have to be looked at again in the light of new statistical and
corpus linguistic techniques. By doing so we will either have even stronger
evidence for the validity of the conclusions, or we will have gained much
more interesting insights into these phenomena.

10

Appendix

Variable levels for contextual predictors:

15

Realization of the subject:

DROP: No explicit subject (pro-drop).

NP: A simple noun phrase.

NULL: No subject (impersonal sentences).

20

PP: A preposition phrase.

PRON: A personal pronoun.

SE: Impersonal sentences with *se*.

Definiteness of the subject/object:

25

ABST: Abstract subjects (mostly verb phrases working as subjects).

DEF: Definite subjects (*el, la, los, etc.*).

IND: Indefinite subjects (*un, una, algunos, etc.*).

NULL: No subject (impersonal sentences).

30

Animacy of the subject/object:

ANI: Animate subjects (human and animal).

INA: Inanimate subjects (including abstract, drop and *se* impersonals).

NULL: No subject (impersonal sentences).

35

Realization of the object:

ADJ: Adjectives and adjective phrases.

ADV: Adverbs and adverbial phrases.

DROP: No object mentioned but implied.

40

NP: Noun phrases.	1
NULL: Verb with no complements.	
PP: Prepositional phrase.	
PRON: Personal pronoun.	
SE: <i>se</i> reflexives.	5
SUB: Subordinate sentences with a complementizer.	
VP: Verbal phrases without a complementizer.	

Sentence type:

	10
ADVER: Adversatives and counterfactual sentences.	
CONDI: Conditional sentences.	
FINAL: Final sentences.	
INDICATIVE: Indicative uses of the subjunctive.	
POTEN: Sentences expressing possibility or doubt, also the default level and prototypical use of the subjunctive.	15
TEMP: Sentences expressing temporal relations.	





POS tags:

	20
A=adjective, C=conjunction, D=determiner, R=adverb, N=noun, V=verb, P=pronoun, S=adposition, F=punctuation mark, W=date, Z=numeral.	

Acknowledgment: Thanks to Doris Schönefeld for her very helpful comments and corrections, the members of the IGRA graduate school for their useful feedback, the participants of the CSDL 2014 Santa Barbara, and the two anonymous reviewers of this article. Usual disclaimers apply.

References

- Academia Española, Real. 2011. Crea: Corpus de referencia del español actual (accessed 30 August 2014).
- Q3 Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes* 22. MIT press.
- Q4 Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991*, 109–149. Springer.
- Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.
- Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295–328.

- Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nessel. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian linguistics* 37(3). 253–291. 1
- Baayen, R. Harald, Peter Hendrix & Michael Ramscar. 2011a. Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. In *Empirically examining parsimony and redundancy in usage-based models, LSA Workshop, January 2011*. 5
- Baayen, R. Harald, Petar Milin, Dusica Filipovic Durffevic, Peter Hendrix & Marco Marelli. 2011b. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438.
- Beuls, Katrien. 2012. Inflectional patterns as constructions: Spanish verb morphology in fluid construction grammar. *Constructions and Frames* 4(2). 231–252.
- Booij, Geert. 2010a. *Construction morphology*. Oxford: Oxford University Press. 10
- Booij, Geert. 2010b. Construction morphology. *Language and Linguistics Compass* 4(7). 543–555.
- Booij, Geert. 2013. Morphology in construction grammar. In Graeme Hoffmann, Thomas Trousdale (eds.), *The oxford handbook of construction grammar*, 255–273. Oxford: Oxford University Press.
- Breiman, Leo. 2001. Random forests. *Machine learning* 45(1). 5–32. 15
- Q5  Bresnan, Joan, Anna Cueni, Tatiana Nikitina, R. Harald Baayen et al. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, 69–94. Amsterdam: KNAW.
- Q6 Cuervo, Rufino José & Ignacio Ahumada. 1981. *Notas a la gramática de la lengua castellana de don andrés bello*. Instituto Caro y Cuervo.
- DeMello, George. 1993.  se subjunctive: A new look at an old topic. *Hispania* 76(2). 235–244. 20
- Elias, Vanessa, Valentyna Filimonova & Andrea Mojedano. 2014. Prescription vs. praxis: The evolution of Spanish imperfect subjunctive. In *New ways of analyzing variation* 43.
- Evans, Vyvyan. 2009. *How words mean: Lexical concepts, cognitive models, and meaning construction*. London: Oxford University Press.
- Evans, Vyvyan. 2010. Figurative language understanding in LCCM theory. *Cognitive linguistics* 21(4). 601–662. 25
- Gaeta, Livio. 2007. On the double nature of productivity in inflectional morphology. *Morphology* 17(2). 181–205.
- Q7 Gili Gava, Samuel. 1983. *Curso superior de sintaxis española: Vox*. Colton Book Imports.
- Glo cott & Peter Dixon. 2004. Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review* 11(5). 791–806. 30
- Q8 Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. Bloomsbury Publishing. 
- Gries, Stefan Th & Anatol Stefanowitsch. 2004. Extending collocation analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1). 97–129.
- Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674. 35
- Janda, Laura A. 2013. *Cognitive linguistics: The quantitative turn*. Berlin: Mouton de Gruyter.
- Johansson, Tobias. 2011. Hail the impossible: p-values, evidence, and likelihood. *Scandinavian Journal of Psychology* 52(2). 113–125.
- Kapatsinski, Vsevolod. 2010. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology* 1(2). 361–393. 40

- Kapatsinski, Vsevolod. 2013. Towards a de-ranged study of variation: Estimating predictor importance with multimodel inference. <http://pages.uoregon.edu/vkapatsi/Deranged.pdf> (accessed 6 June 2015). 1
- Kapatsinski, Vsevolod. 2014. What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology* 11. 1–41.
- Q9 Kempas, Ilpo. 2011. Sobre la variación en el marco de la libre elección entre cantara y cantase español peninsular. *Moenia* 243–264. 5
- Lenz, Rodolfo. 1920. *La oración y sus partes*. Madrid: Centro de estudios históricos.
- Liaw, Andy & Matthew Wiener. 2002. Classification and regression by random forest. *R News* 2(3). 18–22. <http://CRAN.R-project.org/doc/Rnews/> (accessed 13 April 2014).
- Marcos Marín, Francisco et al. 1992. El corpus oral de referencia de la lengua española contemporánea. *Project Report*. Madrid. <ftp://ftp.llff.uam.es/pub/corpus/oral> (accessed 10 1 December 2012).
- Miller, Ralph R., Robert C. Barnet & Nicholas J. Grahame. 1995. Assessment of the Rescorla-Wagner model. *Psychological bulletin* 117(3). 363.
- O'Neill, Paul. 2014. The morpheme in constructive and abstractive models of morphology. *Morphology* 24(1). 25–70.
- Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the language resources and evaluation conference (LREC 2012)*. Istanbul, Turkey: ELRA. 15
- R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/> (accessed 15 June 2015).
- Rescorla, Robert A. & Allan R. Wagner. 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory* 2. 64–99. 20
- Q10 Rojo, Guillermo. 2008. De nuevo sobre la frecuencia de las formas llegara y llegase. In Jörn und Frädislem Albrecht (ed.), *Heidelberger spätlese. ausgewählte tropfen aus verschiedenen lagen der spanischen sprach-und übersetzungswissenschaft. festschrift anlässlich des*, 70, 161–182. Romanistischer Verlag. 25
- Rojo, Guillermo & Victoria Rozas Vázquez. 2014. Sobre las formas en-ra en el español de galicia. http://gramatica.usc.es/~grojo/En_prensa/Sobre_formas_en_ra.pdf (accessed 17 July 2015).
- Schneider, Nathan. 2010. Computational cognitive morphosemantics: Modeling morphological compositionality in Hebrew verbs with embodied construction grammar. In *Proceedings of the 36th annual meeting of the Berkeley linguistics society*. 30
- Schwenter, Scott. 2013. Strength of priming and the maintenance of variation in the Spanish past subjunctive. NWAV. https://www.academia.edu/4857119/_Strength_of_Priming_and_the_Maintenance_of_Variation_in_the_Spanish_Past_Subjunctive-NWAV_42_2013 (accessed 20 June 2015).
- Q11 Steels, Luc. 2011. *Design patterns in fluid construction grammar*, 11. John Benjamins Publishing.
- Stefanowitsch, Ana & Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. 35
- Wilson, Joseph Michael. 1983. *The-ra and-se verb forms in Mexico: A diachronic examination from non-literary sources*. UMass dissertation. <http://scholarworks.umass.edu/dissertations/AAI8401113/> (accessed 10 June 2015).
- Zeileis, Achim, Torsten Hothorn & Kurt Hornik. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17(2). 492–514. 40