The *se-ra* alternation in Spanish subjunctive

Matías Guzmán Naranjo

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 speaker's 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 collostructional analysis I also show that there are repelled and attracted collexemes that distinguish and relate both forms.

1. Introduction

The morphological alternation between *-se* and *-ra* in the Spanish *imperfecto del subjuntivo* ('imperfect subjunctive') has been studied for long but it is still poorly understood, and it remains a challenging problem. We can see the alternation in (1):

(1) a. Si yo fuera 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.'

b. Si yo fuese 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.'

Topics at InfL, 91–120
A. Assmann, S. Bank, D. Georgi, T. Klein, P. Weisser & E. Zimmermann (eds.)
Linguistische Arbeits Berichte 92, Universität Leipzig 2014

Both forms are, at least in principle, possible with all Spanish verbs, and there is no categorical distinction in their use. The difference between both 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 different dialects differ in the attested proportions of use (see section 2), but very 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 speakers' choice, that is, what factors are correlated with the use of *-se* or *-ra*, the emergent patterns 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 for now issues regarding dialectal variation.

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 (2010*a*), 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 informs about the distribution of *-se* and *-ra* in the corpus studied. Section 6 presents a Naive Discriminative learning model that show how different morpho-syntactic and discourse properties of the context correlate with *-se* and *-ra*. Section 7 reports on a collostructional analysis for both forms, and what collexemes can tell us about the semantics of the construction. I discuss the results in section 8, and offer some final remarks in section 9.

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

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 inter-speaker 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.

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[1874]), 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. Cuervo and Ahumada (1981[1874]) noted already in 1874 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 use *-se* almost exclusively, and that this form is almost absent in casual speech in America. Cuervo and Ahumada also claim that *-se* was used in Colombia mainly by writers that 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 a basically 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 non existent, while the use of *-ra* has become widespread.

Gili Gaya (1983: 180-181) also observe that there are regional and personal preferences in the use of *-ra* and *-se*. He also 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 cites Lenz (1920), who claims that when one of the two forms is predominant in use 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 great dialectal variation, 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 it is by no means dead. As for the indicative use of both forms (*el equipo que perdie-ra/se el día de ayer* 'The team that lost.IMP.SUBJ yesterday'), DeMello argues that already around 1950 its use was affected and only present in pedantic writers. The only exceptions seem to be Argentinian Spanish, where it still seems to be somewhat common, and Chilean and Cuban Spanish, where it is occasionally found.

These studies, with the exception of DeMello's, were all done with impressionistic data, and most of them relied solely on the author's intuition of what the distribution of the forms were. 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, the only study of the *-se/-ra* alternation that makes use of quantitative corpus linguistic methods is Schwenter (2013), who claims to have found some effects of PERSON and NUMBER on the choice of the morpheme. Schwenter looks at a large amount of examples[1] from different countries in the CREA corpus (Real Academia Española 2011) and fits 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. 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 (if any) it can correctly predict. It is therefore not possible to contrast his results with those of the present study in any meaningful way.

In summary, most studies done 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.

3. The imperfect subjunctive construction

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

[1]However, an important shortcoming of Schwenter's study is that he only considered 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.

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

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

What (2) says is basically that there is a semi-abstract imperfect subjunctive construction which instantiates 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 value[3] not derivable from either the morpheme nor the verb. In this analysis both *-se* and *-ra* are more specific constructions that instantiate the more general abstract construction in (2) and have the forms in (3):

(3) a. $[[X_{vi}]\ \text{-}ra_j]_v \leftrightarrow [SEM_i$ in imperfect tense subjunctive + $PRAG_l$ + $PRAG_j]$
b. $[[X_{vi}]\ \text{-}se_k]_v \leftrightarrow [SEM_i$ in imperfect tense subjunctive + $PRAG_l$ + $PRAG_k]$

What this means is that both constructions *-se* and *-ra* instantiate the same grammatical core construction in (2), retain the pragmatic value associated with it ($PRAG_l$) but specify additional pragmatic information associated exclusively to the specific form in question ($PRAG_j$ and $PRAG_k$). This analysis captures well the fact that both constructions have indeed the same grammatical function, but that there seem to be important differences between both forms. The interesting issue thus is to investigate what $PRAG_l$, $PRAG_j$ 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 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 levels that deal independently with TAM and person and number. This representation 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 Evans (2009, 2010).

Analyzing inflectional morphology from a construction grammar perspectives is, as far as I am aware, not common practice. A notable exception can be seen in Beuls (2012), who within the framework of Fluid Construction Grammar (Steels 2011), developed a full implementation of Spanish inflectional morphology (see also Schneider 2010). She does not address the issue of this particular alternation, though.

4. Material

The corpus used for this study was the Corpus Oral de Referencia de la Lengua Española Contemporánea, CORLEC, (Marcos Marín et al. 1992). The CORLEC 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, formating 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 (utterance) boundaries. This results in a division that is a product of 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 cases 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 set of sentences (plus all the occurrences for *-se*) some cases were removed if they were clear errors or instances of a different genre, e.g. people reading poetry. After all these fixes, the total number of *-ra* sentence was 184 and for *-se* 183.

Besides the study by Schwenter (2013), there are no proposals in the literature for any particular set of variables that could influence the *-se/-ra* alternation. Because of this, I include in this study also variables that have been found to be relevant in distinguishing other alternations, even if there seems to be no reason for including them for analyzing a morphological alternation. The

variables related to the verb that appears in the imperfect subjunctive form are the following. The variable VERB is simply which verb in question was used in the imperfect subjunctive, which should tell us whether there are lexical preference in the alternation. Directly related variables, and also suggested by Schwenter[4], are PERSON and NUMBER of the verb in imperfect subjunctive. Also closely related to the variable VERB is the verb ending (often referred to in the literature as thematic vowel of the verb) *-ar, -er* or *-ir* (in the models coded as TYPE). This variable could be important for priming reasons (the vowel /a/ could prime *-ra* and /e/ could prime *-se*). Additionally, an interesting variable that could have an effect is whether the verb appearing in imperfect subjunctive has a modal meaning (coded as MODAL). The status of modals in Spanish is not without debate, for simplicity I took the verbs *querer* 'want', *poder* 'can', *deber* 'must', *soler* 'do often', *tener* 'have (to)' to be modals. The main reasons for including these verbs as modals is that they 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 7 this decision seems to be justified. Then, we have lexical variables associated with the choice of verb (VERB, MODAL, TYPE), and grammatical variables (number and person).

A different set of variables is recruited from elements related to the grammatical and discourse context that the verb appears in. I coded all *-se* sentences and the randomly chosen *-ra* sentences for: ANIMACY OF THE SUBJECT (`NP`, `pronoun`, `drop`, `null`, etc.), DEFINITENESS OF THE SUBJECT[5], REALIZATION OF THE SUBJECT, ANIMACY OF THE OBJECT[6], DEFINITENESS OF THE OBJECT, REALIZATION OF THE OBJECT (`NP`, `PP`, `pronoun`, `null`, etc.), and TYPE OF SENTENCE. For this final variable the following types were considered: `conditional` (expressing a condition on which something happens), `final` (expressing desire or determination that something happens), `indicative` (indicative use of the subjunctive), `temporal` (expressing temporal relations), `adversative`

[4]Schwenter's proposal for considering 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. It also seems trivially true that any form will prime itself.

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

[6]I also considered in this category 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.

(comparison or opposition to something), and `potential` (other uses where possibility or probability are conveyed by the subjunctive). Related to the type of sentence is whether the words *que* 'that.COMPL' and *si* 'if' introduced 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 other contextual variables I included were CATEGORY OF THE NEXT WORD and CATEGORY OF THE PRECEDING WORD, these were extracted form the first letter in the POS tags provided by FreeLing plus a `X` category 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. Distribution of the alternation

After removing cases with incomplete information and some clear errors in the extraction, the total number of observations was 1269, 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 1.

	se	ra	total
total cases	191	1078	1269
total sentences	171	911	1081
total verbs	97	228	325

Table 1: Total number of occurrences, sentences and verbs with the forms *-se* and *-ra*

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[7]. The morpheme *-ra* seems to appear with modals and verbs ending in *-er* more

[7] 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.

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*, but, as we will see in the collostructional analysis, there are reasons to suspect that their overlap is coincidental.

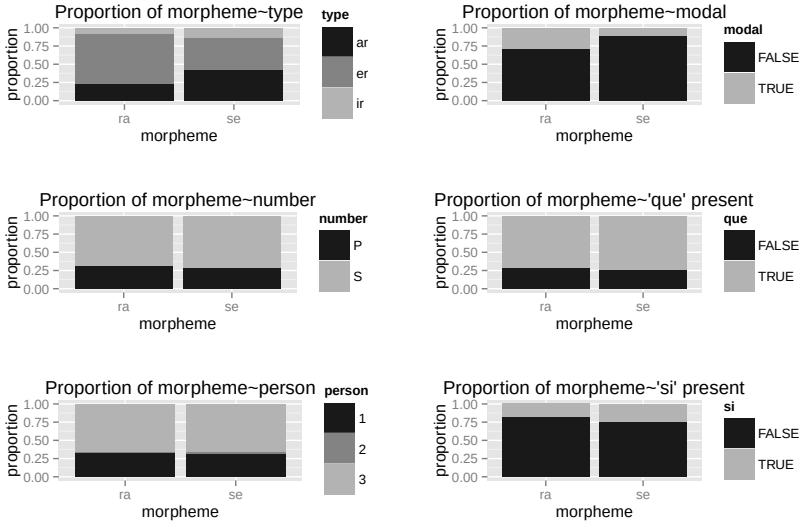

Figure 1: Proportions of the variables TYPE, PERSON, NUMBER, MODAL, QUE, SI for *-se* and *-ra*.

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 difference in object phrases, but it is apparent that *-ra* appears with more sentences without overt subjects than *-se*. For objects, the differences are larger. The form *-se* prefers noun phrases, while *-ra* shows almost the same preference for noun phrases and verb phrases.

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). We can notice 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.

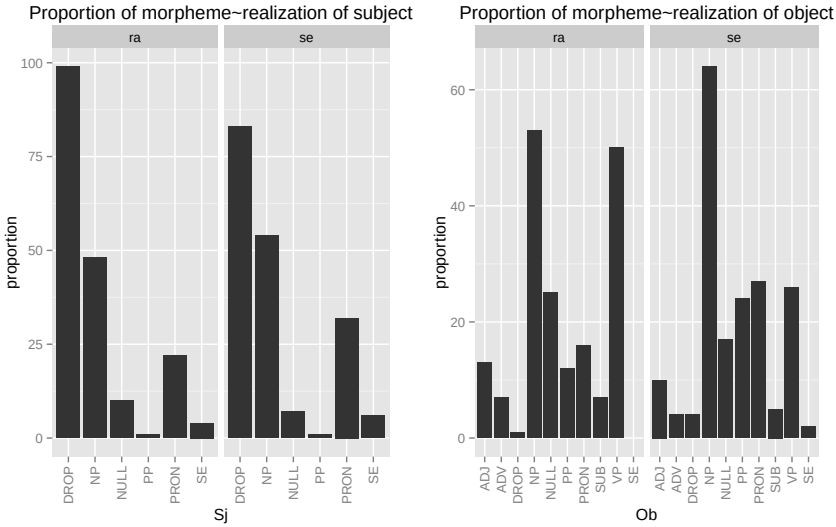

Figure 2: Proportions of the variables REALIZATION OF SUBJECT and REALIZATION OF OBJECT for *-se* and *-ra*. DROP = no overt subject near the verb, NP = noun phrase (with or without determiner), NULL = impersonal uses like existential *haber*, PRON = single pronoun (also relatives, demonstratives and 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.

In Figure 4 and Figure 5 we can see the distributions of the grammatical categories of the preceding and following words[8].

From both figures, we can see that there does not seem to be much difference when it comes to the preceding grammatical category between both morphemes. The following grammatical category does show some differences, 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 next section.

The next factor considered was LENGTH OF SENTENCE. As mentioned above,

[8] For a full list of what each POS tag means see http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html.

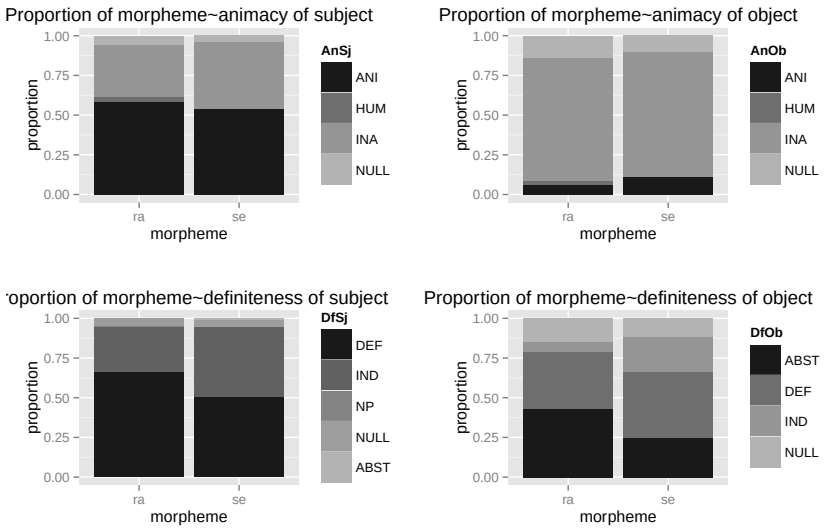

Figure 3: Proportions of the variables ANIMACY OF SUBJECT, ANIMACY OF OBJECT, DEFINITENESS OF SUBJECT, DEFINITENESS OF OBJECT for *-se* and *-ra*. NULL = no subject or object, ABST = for phrases different from NPs that work as the subject or first complement of the verb. DEF and IND are definite and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

there are repeated sentences in the data for the cases where a single sentence contains more than one case of the construction. Figure 6 and 7 shows the distribution of the length of sentence for each form considering only the manually coded cases, including repeated sentences.

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, unexpectedly, *-se* also appears with some verb types that do not appear with *-ra*. The individual lists of unique verbs appearing with either *-se* or *-ra* are shown in Tables 2 and 3 respectively.

A detailed analysis of collexemes will be presented in section 7, but both these tables already suggest that there are lexical preferences associated with both morphemes. We can see that *deber* ('must'), a modal verb, never appears

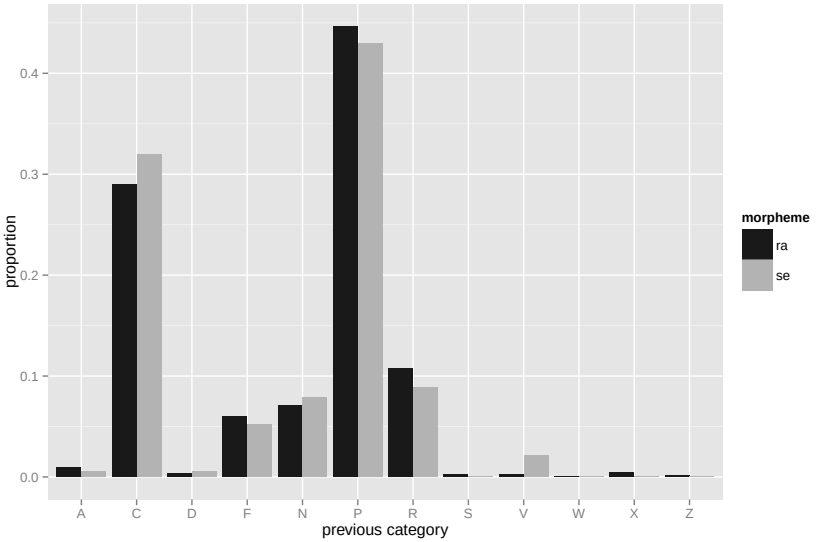

Figure 4: Proportion of preceding grammatical category present for *-se* and *-ra*.

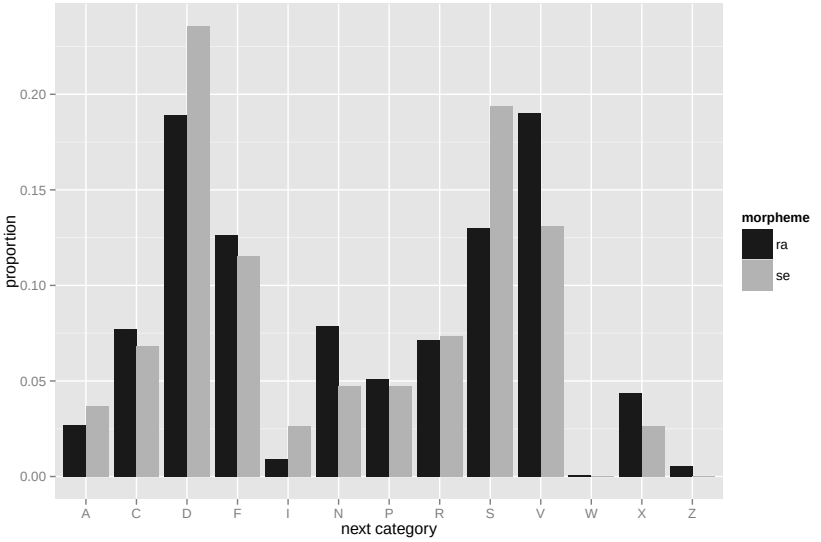

Figure 5: Proportion of next grammatical category present for *-se* and *-ra*.

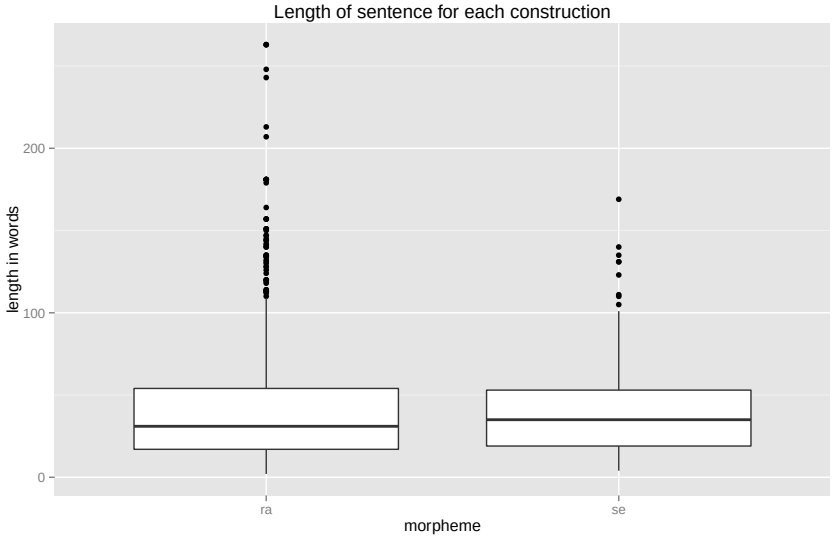

Figure 6: Length of sentence for *-se* and *-ra*.

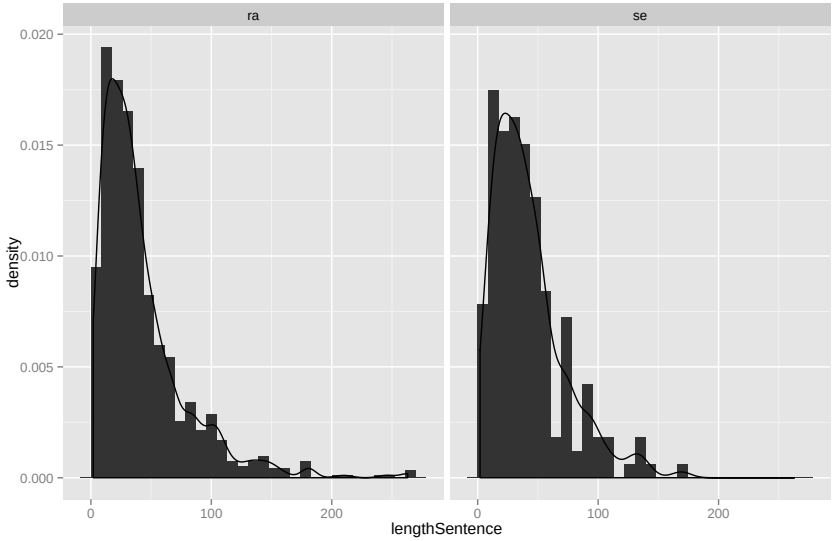

Figure 7: Histogram of length of sentence for *-se* and *-ra*.

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
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

Table 2: Verbs that appear with *-se* but not with *-ra*

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
seguir	follow	4	0.00371058
aparecer	appear	3	0.00278293
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
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
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
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
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
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

Table 3: Most frequent verbs that appear with *-ra* but not with *-se*

with *-se*, suggesting that modality of the verb might play a role in distinguishing both forms. 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 this method 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.

6. 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 or AIC scores. 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 or AIC scores, instead of using these method I will focus mostly on the C score of the model for model selection.

6.1. 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), and the realization of subject (SJ) and object (OB).

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, Baayen, Milin, Đurđević, Hendrix and Marelli 2011, Baayen 2011, Baayen, Hendrix and Ramscar 2011, Baayen et al. 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 et al. 1972), which has proven to be an excellent model for animal and human

learning (Miller et al. 1995). 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).

For model assessment I will mainly use the Area Under the Roc Curve value (C). The C score can go 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 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.

6.2. Morpho-syntactic and discourse factors

The smallest model that best fits the data has the following formula: `morpheme ~ modal + DfSj + DfOb + sentenceType + lengthSentence + 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` were not significant for the model (i.e. they did not significantly improve the overall C score of the model). The confusion matrix for this model is shown in Table 4. We can see that the model fits the data very well, with very few errors.

Confusion Matrix		
	Prediction	
Reference	ra	se
ra	160	24
se	32	151
Accuracy: 0.8474%		
C score: 0.9174		

Table 4: Confusion Matrix for Model A.

By far, the strongest predictor was VERB, which suggests very strong lexical preferences in the construction. Since section 7 will deal exclusively with the issue of lexical effects, I will not go into a detailed discussion of this predictor here. The best individual predictors for *-ra* and *-se* are given in Figures 8 and 9:

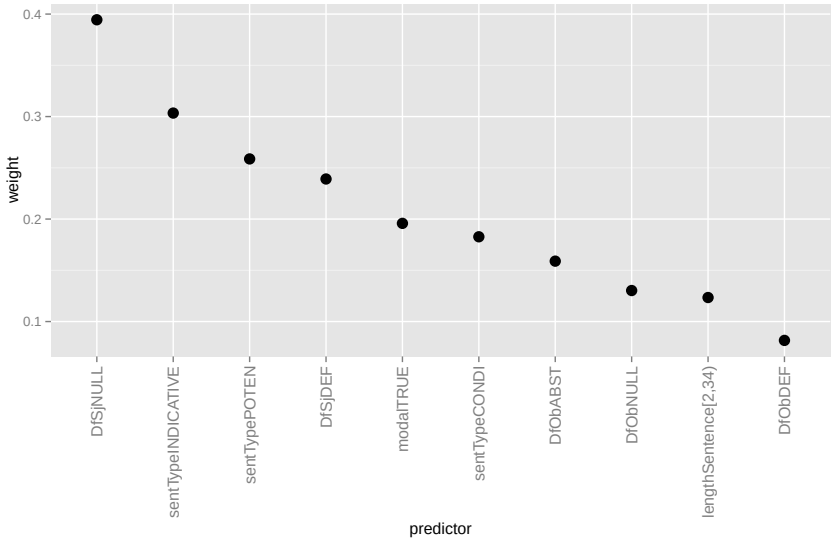

Figure 8: Best 10 predictors for *-ra*

From these figures we can see that most predictors other than VERB are, for the most part, relatively weak. The strongest predictor for both *-se* and *-ra* was DEFINITENESS OF THE SUBJECT, with null subjects predictive of *-ra* and

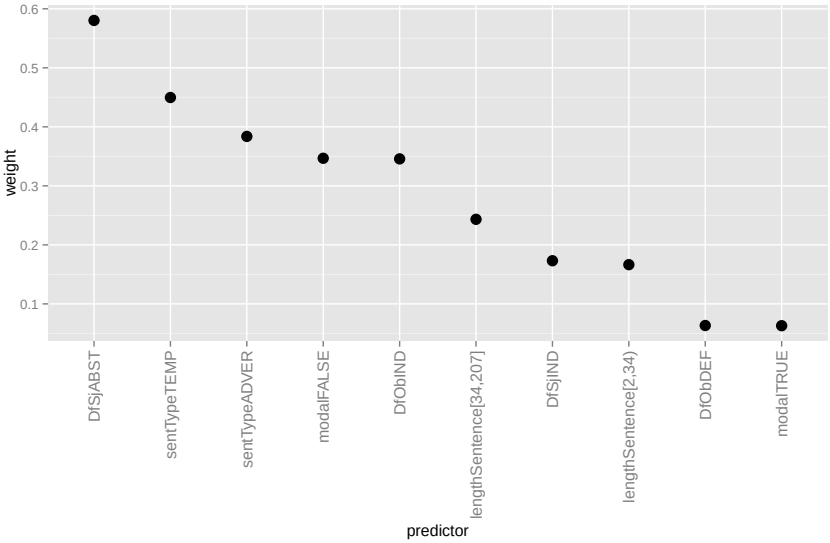

Figure 9: Best 10 predictors for *-se*

abstract subjects (those different from NPs) predictive of *-se*. We see, as second best predictor of both, the type of sentence, with adversatives and temporal sentences predictive of *-se* and indicative 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. Finally *-se* seems to be preferred over *-ra* longer sentences, which might be indicative of discourse issues like formality or type of turn (e.g. monologue vs conversation).

A detailed interpretation for each single level of each predictor is not easy (and because of their low scores not very enlightening), but from this discussion it is clear that the strongest predictors of both *-se* and *-ra* are not grammatical levels on the verb, but elements of discourse and context in the sentence. This contradicts the results by Schwenter (2013).

6.3. Model evaluation and overfitting

Although the previous model achieved high accuracy, it is important to evaluate how much the patterns observed are specific to this particular data-set, and

which are more likely to be part of the alternation as a whole. There are two techniques we can use to test this. The first one is bootstrapping the model by splitting the data into training and testing portions, and repeating the process multiple times (30 in this case), and the second one is using machine learning algorithms that are less prone to overfitting[9]. The results from bootstrapping Model A are presented in Table 5.

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		

Table 5: Mean Confusion Matrix for bootstrap of Model A.

We can see in Table 5 that there is a significant drop in accuracy and C score, but nevertheless the model seems to retain a predictive capability well above random chance. This gives us some confidence that the model is on the right track, despite not being as powerful as initially thought.

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 does splitting of the data during training. Fitting the same model but with a Random Forest classifier we get the results in Table 6.

From Table 6 we can see again that there is a considerable reduction in accuracy, but nonetheless the model still performs above chance. From the results of both evaluations we can conclude that the observed correlations are in fact real, but that the original model is overfitting the data to a certain extent. Also interesting is that if we examine the importance of each predictor as given by Random Forest (Table 7), we get a very similar picture to that in the Naive Discriminative Learning model. The strongest predictors in Random Forest were DEFINITENESS OF SUBJECT and DEFINITENESS OF OBJECT, followed by the

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

Confusion Matrix		
	Prediction	
Reference	ra	se
ra	134	50
se	72	111
Accuracy: 0.6676%		
C score: 0.7239		

Table 6: Confusion Matrix for Random Forest fit of Model A.

verbs *poder*, *querer*, and *cambiar*, and MODAL, which is roughly similar to what we saw with the NDL classifier.

Predictor	-ra	-se	MeanDecrease Accuracy	MeanDecrease Gini
DfOb	12.3634078	18.49193419	19.7888637	10.99018513
DfSj	11.4994495	11.38458302	14.1354551	6.13536258
verb *poder*	6.7802031	11.81506783	11.9360860	2.91252338
modal	5.2346876	12.49406023	11.7691853	3.92447546
verb *querer*	6.3810543	11.29882105	10.9612298	3.01719544
sentType	2.9030475	10.98845199	9.5821425	7.45687579
verb *cambiar*	7.9014889	5.37849872	8.5969787	1.17386609
lengthSentence	7.8189127	5.38468293	8.3552875	12.13443005

Table 7: Best predictors for Random Forest fit of Model A.

We can conclude that although our model is not a perfect fit, there are contextual factors, as well as lexical effects of the verb, that are correlated with the forms in the *-se/-ra* alternation.

7. Collostructional analysis

Finally, to investigate in depth the lexical preference 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 also looked at the whole data-set for this part of the analysis.

7.1. Attracted collexemes

First we look at the 20 collexemes that are most strongly attracted to both *-ra* (Table 8) and *-se* (Table 9). The first interesting fact that can be observed is that the top three positions for *-ra* are occupied by verbs that can be typically used as modals: *querer* 'want', *poder* 'can' and *deber* 'must'[10]. 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 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 that the actual number of cooccurences 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.

7.2. 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 10) and *-se* (Table 11). The first interesting thing we find 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*. 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,

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

N	Verb	Gloss	Co-occurrences	Expected Frequency	Observed Frequency	Fisher's p
1	querer	want	114	12.164734	2375	1.954e-65
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
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
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	constituir	constitute	2	0.174155	31	1.481e-02

Table 8: First 20 attracted collexemes for *-ra*.

N	Verb	Gloss	Co-occurrences	Expected Frequency	Observed Frequency	Fisher's p
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

Table 9: First 18 attracted collexemes for *-se*.

which suggests that it is in fact the periphrastic future form that 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.

N	Verb	Gloss	Co-occurrences	Expected Frequency	Observed Frequency	Fisher's p
1	ir	go	5	45.435	7592	1.553e-13
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
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
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	producir	produce	1	2.086	348	7.287e-01

Table 10: First 20 repelled collexemes for *-ra*.

Another apparent pattern we can 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.

For *-se* there are no repelled lexemes that reach significance ($p<0.05$). In this cases *ir* is probably related to the same issues discussed for *-ra*, but an

N	Verb	Gloss	Co-occurrences	Expected Frequency	Observed 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	derivar	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

Table 11: First 20 repelled collexemes for *-se*.

interpretation for *hacer* is less clear. The fact that all p-values are too large, and that the difference between observed co-occurence and expected co-occurence is too small means that it is quite possible that the distribution of most lexemes presented in Table 11 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. More interesting yet 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.

7.3. Contrastive collexemes

We can also contrast the collexemes for *-se* and *-ra* by evaluating whether the proportion observed for each verb for each form is likely due to chance as it would be expected from the proportion of both constructions, or if there is likely to be a preference. This method is simply testing the null hypothesis 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 12.

N	Verb	Gloss	-ra	-se	Global Observed 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

Table 12: Contrastive collexemes for *-se* and *-ra*.

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 favour 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 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*.

8. Discussion

As we saw, the Naive Discriminative Learning model showed 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 not. These effects remained present and significant after controlling for overfitting. The model also presented evidence for strong lexical effects, both in

the lexical choices of individual verbs, as well as 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'), 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 most differentiate both constructions are *querer* ('want') and *poder* ('can'). All these facts very strongly support the case for discourse difference between both forms, but also for some discourse similarities.

These results are very 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. Especially interesting is that neither NUMBER nor PERSON helped distinguish between both forms. This can be understood withing 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 are 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 (4) and (5):

(4) $[[X_{vi}]\ \text{-}Y_{se/ra}\ [\text{PERSON}]\ [\text{NUMBER}]]_v \leftrightarrow [SEM_i$ in imperfect tense subjunctive + $PRAG_l]$

Based on the results of the models we can propose that the NUMBER and PERSON constructions are instantiated on the abstract construction in (4). This means that at the level of (4) both NUMBER and PERSON are free slots in the constructions. The more specific constructions for *-se* and *-ra* would be the following:

(5) a. $[A_{vi(j)}\ \text{-}ra_j + \text{PERSON/NUMBER}]_v \leftrightarrow [SEM_i$ in imperfect tense subjunctive + $PRAG_l + PRAG_j]$
b. $[B_{vi(k)}\ \text{-}se_k + \text{PERSON/NUMBER}]_v \leftrightarrow [SEM_i$ in imperfect tense subjunctive + $PRAG_l + PRAG_k]$

Where A and B stand for concrete lexical choices (not free slots as before) 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 type of sentence where the subjunctive appeared. At this level both PERSON and NUMBER are not free slots, but are inherited from the more abstract construction in (4) (and the individual constructions for number and person). For $PRAG_l$, discourse preferences common to both forms, it was not 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 more work is required. It is possible that this variable is only acting as proxy for some semantic effect.

Independently of whether this specific analysis is correct, the results of the models and the collostructional analysis are empirical evidence that lend some support of 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 can only be explained from a usage-based perspective where each exemplar counts, and each exemplar can be richly represented including the context it appeared in.

9. 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 is most likely wrong. However, it must be emphasized that the models presented 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. In order to explore these issues an actual implementation in Fluid Construction Grammar would be necessary.

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 the 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 modelling 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 (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 in principle impossible to ever know whether the statistical model we chose reached ceiling or whether there are other still 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 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 conclusions, or we will have gained much more interesting insights into these phenomena.

References

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 and Tore Nesset (2013): 'Making choices in Russian: pros and cons of statistical methods for rival forms', *Russian linguistics* 37(3), 253–291.

Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix and Marco

Marelli (2011): 'An amorphous model for morphological processing in visual comprehension based on naive discriminative learning.', *Psychological review* **118**(3), 438.

Baayen, R. Harald, Peter Hendrix and Michael Ramscar (2011): Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. In: *Empirically examining parsimony and redundancy in usage-based models, LSA workshop*. .

Beuls, Katrien (2012): 'Inflectional patterns as constructions: Spanish verb morphology in Fluid Construction Grammar', *Constructions and Frames* **4**(2), 231–252.

Booij, Geert (2010*a*): *Construction morphology*. Oxford: Oxford University Press.

Booij, Geert (2010*b*): 'Construction morphology', *Language and Linguistics Compass* **4**(7), 543–555.

Booij, Geert (2013): 'Morphology in construction grammar', *The Oxford handbook of Construction Grammar* pp. 255–273.

Breiman, Leo (2001): 'Random Forests', *Mach. Learn.* **45**(1), 5–32.
URL: *http://dx.doi.org/10.1023/A:1010933404324*

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, R. Harald Baayen et al. (2007): 'Predicting the dative alternation', *Cognitive foundations of interpretation* pp. 69–94.

Cuervo, Rufino José and Ignacio Ahumada (1981[1874]): *Notas a la Gramática de la lengua castellana de don Andrés Bello*. Instituto Caro y Cuervo.

DeMello, George (1993): '-Ra vs.-se subjunctive: a new look at an old topic', *Hispania* pp. 235–244.

Evans, Vyvyan (2009): *How words mean: lexical concepts, cognitive models, and meaning construction*. Oxford University Press London.

Evans, Vyvyan (2010): 'Figurative language understanding in LCCM theory', *Cognitive linguistics* **21**(4), 601–662.

Gili Gaya, Samuel (1983): *Curso superior de sintaxis española: Vox*. Colton Book Imports.

Gries, Stefan Th. and Anatol Stefanowitsch (2004): 'Extending collostructional analysis: A corpus-based perspective onalternations", *International journal of corpus linguistics* **9**(1), 97–129.

Gries, Stefan Thomas (2003): *Multifactorial analysis in corpus linguistics: A study of particle placement*. Bloomsbury Publishing.

Janda, Laura A. (2013): *Cognitive Linguistics: The Quantitative Turn*. Mouton de Gruyter, Berlin.

Kapatsinski, Vsevolod (2010): 'Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology', *Laboratory Phonology* **1**(2), 361–393.

Kapatsinski, Vsevolod (2014): What is grammar like? A usage-based constructionist perspective. *In:* B. Cartoni, D. Bernhard and D. Tribout, eds, *Theoretical and computational approaches to morphology*. .

Lenz, Rodolfo (1920): *La oración y sus partes*. Number 5, Revista de filología española.

Liaw, Andy and Matthew Wiener (2002): 'Classification and Regression by randomForest', *R News* **2**(3), 18–22.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Marcos Marín, Francisco et al. (1992): 'El Corpus Oral de Referencia de la Lengua Española contemporánea', *Project Report. Madrid. Publisher in ftp://ftp. lllf. uam. es/pub/corpus/oral* .

Miller, Ralph R., Robert C. Barnet and Nicholas J. Grahame (1995): 'Assessment of the Rescorla-Wagner model.', *Psychological bulletin* **117**(3), 363.

Padró, Lluís and Evgeny Stanilovsky (2012): FreeLing 3.0: Towards Wider Multilinguality. In: *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. ELRA, Istanbul, Turkey.

R Core Team (2014): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Real Academia Española (2011): 'CREA: Corpus de referencia del español actual'.

Rescorla, Robert A., Allan R. Wagner et al. (1972): 'A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement', *Classical conditioning II: Current research and theory* **2**, 64–99.

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*. .

Schwenter, Scott (2013): Strength of Priming and the Maintenace of Variation in the Spanish Past Subjunctive. NWAV.
URL: *https://www.academia.edu/4857119/_Strength_of_Priming_and_the_-Maintenance_of_Variation_in_the_Spanish_Past_SubjunctiveNWAV_42_2013*

Steels, Luc (2011): *Design patterns in fluid construction grammar*. Vol. 11, John Benjamins Publishing.

Stefanowitsch, Anatol and Stefan Th. Gries (2003): 'Collostructions: Investigating the interaction of words and constructions', *International journal of corpus linguistics* **8**(2), 209–243.

Wilson, Joseph Michael (1983): 'The-ra and-se verb forms in Mexico: a diachronic examination from non-literary sources'.