

# Lectal contamination

## Evidence from corpora and from agent-based simulation

Dirk Pijpops

RU Lilith, University of Liège

This paper presents evidence from both corpora and agent-based simulation for the effect of lectal contamination. By doing so, it shows how agent-based simulation can be used as a complementary technique to corpus research in the study of language variation. Lectal contamination is an effect whereby the words that are typical of a language variety more often appear in a morphosyntactic variant typical of that same variety, even among language use from a different variety. This study looks at the Dutch partitive genitive construction, which exhibits variation between a “Netherlandic” variant with *-s* ending and a “Belgian” variant without *-s* ending. It is shown that the probability of the Belgian variant without *-s* increases among more “Belgian” words, in the language use of both Belgians and people from the Netherlands. Meanwhile, an agent-based simulation reveals the crucial theoretical preconditions that lead to this effect.

**Keywords:** agent-based modelling, simulation, lectal contamination, mixed regression modelling, partitive genitive

### 1. Introduction

The present paper has a twofold goal. First, it sets out to investigate lectal contamination. This an effect that may take place when two language varieties, e.g. two dia-, socio- or regiolects, have a different incidence of the variants of a morphosyntactic alternation. One example of such an alternation is the presence vs. absence of *of* in Example (1), where the variant with *of* occurs more often

in American English, while the variant without *of* is relatively more popular in British English (Algeo, 2006: 64).<sup>1</sup>

- (1) *Both (of) these names* refer to its farm food value. (taken from Algeo, 2006: 64)

‘Lectal contamination’ then occurs when the words that are more often used in one language variety show a bias towards the morphosyntactic variant typical of that same variety, even among speakers of another variety. For example, Britons would more often use the “American” variant with *of* when using “American” words than when using “British” words.

Another example can be found in the Dutch partitive genitive construction. This construction appears in a morphological variant with an *-s* ending, as in Example (2), and a variant without an *-s* ending, as in Example (3).<sup>2</sup> The variant with *-s* ending is more prevalent in the Netherlandic variety of Dutch, while the variant without *-s* maintains a stronger presence in the Belgian variety (Broekhuis, 2013: 426; van der Horst, 2008: 1624–1625).<sup>3</sup> Still, both variants commonly occur throughout either variety (Pijpops & Van de Velde, 2015, 2018). Lectal contamination would then predict that, within the Belgian variety, lexemes that are more popular in the Netherlandic variety would prefer the variant with *-s*, whereas those that more often appear in the Belgian variety would more often exhibit the variant without *-s*. The same would hold within the Netherlandic variety.

- (2) Is er nog *iets leuks* te beleven? (taken from Pijpops & Van de Velde, 2016: 553)  
“Is there still something fun to do?”
- (3) of er hier nog *iets leuk* te beleven valt  
(taken from Pijpops & Van de Velde, 2016: 553)  
“... whether there is still something fun to do here?”

The second goal of this paper is to exemplify how agent-based modelling may be used to support corpus research in the study of language variation. An agent-

1. This example only pertains to *both (of) these*, not *both of the*, *both of my*, etc. The situation is different when other determiners are used (Algeo, 2006: 64).

2. The *-s* ending is a remnant of the historical genitive case from Dutch, hence the name of the construction. From a purely synchronic point of view, the ending is better viewed as an isolated suffix than as a case marker. The partitive genitive construction itself is highly productive in Dutch, both with and without the *-s* ending, and partitive genitives with neologisms, such as *iets klimaatneutraals* “something climate neutral” can be readily found on the internet.

3. For reasons of clarity, I will use the adjective *Netherlandic* when referring to the country of the Netherlands, and reserve the term *Dutch* to talk about the Dutch language. As such, I call the variety of Dutch spoken in the Netherlands ‘Netherlandic Dutch’, and the variety of Dutch spoken in Belgium ‘Belgian Dutch’.

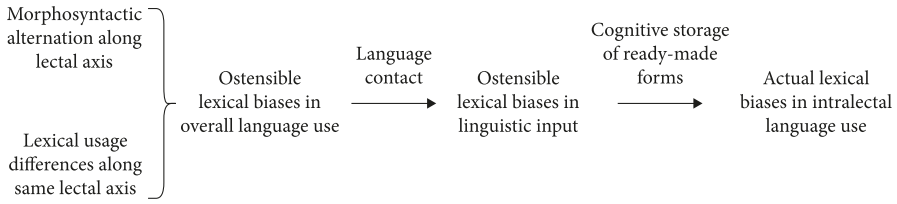
based model is a computer simulation in which various computational ‘agents’ interact with one another according to simple rules defined by the researcher. The intention is to study what kind of tendencies emerge at the community level from these local interactions between agents (Gilbert, 2008; Phan & Varenne, 2010; Steels, 2011). This makes agent-based modelling particularly suited to investigate complex adaptive systems, such as languages (Beckner et al., 2009; Steels, 2000).

Agent-based models are already used extensively in evolutionary linguistics to study the origins of language (e.g. Beuls & Steels, 2013; Jaeger et al., 2009 and references cited therein), and are beginning to find their way into historical linguistics to study language change (e.g. Bloem et al., 2015; Blythe & Croft, 2012; Pijpops et al., 2015). Still, their use in variational linguistics has been more limited (for exceptions, see Fagyal et al., 2010; Karjus & Ehala, 2018). Moreover, the available linguistic studies that employ agent-based modelling only rarely combine the technique with corpus research. This is regrettable, as both methodologies are highly complementary, as the present paper hopes to show.

Section 2 discusses the preconditions for lectal contamination to take place, as well as the predictions that it makes. Section 3 then presents a corpus investigation that tests the effect of lectal contamination, using the Dutch partitive genitive alternation as a case study. Data are drawn from the Corpus of Spoken Dutch (Oostdijk et al., 2002; van Eerten, 2007) and the Sonar corpus of written Dutch (Oostdijk et al., 2013b). Next, Section 4 reports on an agent-based model that implements the mechanism of lectal contamination *in silico* and ties its results back to the findings of the corpus study. Finally, Section 5 summarizes the conclusions.

## 2. Preconditions and predictions of lectal contamination

The mechanism theorized to cause lectal contamination is visualized in Figure 1. Lectal contamination may occur when a morphosyntactic alternation is stratified along a lectal axis, e.g. a distinction between regio-, dia-, socio- or ethnolects. Two potential examples were already mentioned in Examples (1) to (3). Another could be the pluralization of presentational *haber* in Spanish, where the singular form in Example (4) functions as the standard variant, while the plural form in Example (5) is especially popular in Latin American varieties (Bentivoglio & Sedano, 2011; Claes, 2015; Pérez-Martín, 2007). A fourth possible example is the choice between the German suffixes *-ation* and *-ung* and in Examples (6) and (7). The *-ung* suffix is dominant in Germany and Austria, while the *ation* variant maintains a strong foothold in Swiss German (Dürscheid et al., 2018). Again, both variants do commonly occur throughout the entire German language area, however.



**Figure 1.** Flowchart of the mechanism of lectal contamination

- (4) Y de, por eso siempre *va a haber* personas subyugadas y *va a haber* pobres y *va a haber* ricos. (taken from Claes, 2015: 2)  
 “And of, therefore, there will always be oppressed people and there will be poor and there will be rich.”
- (5) Y, e, *han habido* ciertos cambios en, en la sociedad. (taken from Claes, 2015: 2)  
 “And, e, there have been certain changes in, in society.”
- (6) Es besteht ein limitiertes Platzangebot, frühzeitige *Reservierung* wird empfohlen. (taken from Dürscheid et al., 2018)  
 “Seats are limited, early reservation is recommended.”
- (7) Wegen der hohen Nachfrage empfiehlt die EVB eine frühzeitige *Reservierung*. (taken from Dürscheid et al., 2018)  
 “Because of high demand, the EVB recommends early reservation.”

The same lectal axes often also exhibit differences in their usage of lexical items. For instance, the noun *backpack* is fairly frequent in the one billion Corpus of Contemporary American English (6325 occurrences, COCA, Davies, 2008–), but less so in the 100 million British National Corpus (35 occurrences, BNC, Davies, 2004). Meanwhile, the noun *rucksack* is often used in the BNC (237 occurrences), but comparatively less so in the COCA (416 occurrences). Such differences pervade the lexicon, often without language users being aware of them (Ruetten et al., 2016; Wieling & Nerbonne, 2015).

As a result, it would appear as if there are lexical biases when looking at overall language use: naturally, morphosyntactic variants typical for lect A would be found more often among lexical items typical for lect A. Put concretely, when looking at English language use from both the United States and Britain, the variant with *of* would be expected to be more dominant for *both (of) these backpacks* than for *both (of) these rucksacks*. The reason is simply that a greater percentage of the instances of the *both (of) these backpacks* would stem from Americans. In other words, it would appear at first sight as if the noun *backpack* has a lexical bias for the variant with *of*, whereas the noun *rucksack* has a lexical bias for the variant

without *of*. Of course, if the American data and the British data would be studied separately, such an ostensible lexical bias would normally disappear.

Similarly in Dutch, partitive genitives with typically Netherlandic lexical material such as *iets bijzonder(s)* “something special” would be expected to exhibit an ostensible lexical bias for the “Netherlandic” variant with *-s*. Meanwhile, partitive genitives with typically Belgian lexical material such as *iets speciaal(s)* “something special”, would more often appear in the “Belgian” variant without *-s*. Again, this ostensible lexical bias would be expected to disappear once Netherlandic and Belgian Dutch are studied separately.

Now, individual language users are often exposed to language use from various lects, either through direct contact with language users that speak a different lect or via all sorts of traditional or social media. As such, there would be ostensible lexical biases present in the language use that individual language users hear and read. In other words, if there is language contact between both lects, then these lexical biases are part and parcel of the linguistic input of language users. For instance, a Sheffielder would be relatively more exposed to the *of*-variant of *both (of) these backpacks* than to the *of*-variant of *both (of) these rucksacks*, or an Antwerpian would hear or read relatively more often the *-s* variant of *iets bijzonder(s)* “something interesting” than the *-s* variant of *iets speciaal(s)* “something special”.

It is one of the key foundations of the theoretical framework of usage-based linguistics that language users are sensitive to the language use around them, and tend to reproduce the tendencies in this language use, even if they are not consciously aware of them (Bybee, 2010, 2013; Diessel, 2015). This alone does not yet entail that language users are receptive to ostensive lexical biases in morphosyntactic variation. For this to be the case, they would need to store phrases such as *both of these backpacks* or *iets bijzonder* “something special” in memory as ready-made forms. Still, there is in fact strong empirical evidence that language users can do so, at least occasionally (Arnon & Snider, 2010; Dąbrowska, 2014; Tremblay et al., 2011). In addition, Diessel (2019: 113–195) discusses theoretical reasons why “filler-slot relations (...) are also influenced by speaker’s and listener’s experience with particular lexemes and constructions” (Diessel, 2019: 195).

The effect of lectal contamination then arises once language users start reproducing these lexical biases. As for the English example, there would be a higher probability that the Sheffielder would produce the *of*-variant for *both (of) these backpacks* than for *both (of) these rucksacks*. The reason is simple: they would have heard the *of*-variant for *both (of) backpacks* comparatively more often than the *of*-variant for *both (of) these rucksacks*. Turning to the Dutch example, one would expect a higher probability that the Antwerpian produces the *-s* variant of *iets bijzonder(s)* “something special” than of *iets speciaal(s)* “something special”.

The reason is the same: they would have heard *iets bijzonder(s)* comparatively more often with the *-s* ending.

To continue with both examples, the lexical biases are now produced by individual Britons and individual Belgians. In other words, the biases have penetrated or “contaminated” a single lect. As a result, the biases would still be present in the results if the analysis distinguishes between utterances from Americans and Britons or between utterances from Belgians and people from the Netherlands. Lectal contamination is at work when such intralectal lexical biases have emerged. In sum, the following four preconditions are theoretically required for lectal contamination to take place.

- i. A morphosyntactic alternation, with one variant being more strongly present in one lect than in another lect
- ii. Probabilistic differences in lexical usage between both lects
- iii. Language contact between both lects
- iv. The cognitive storage of ready-made forms by language users

To be clear, the following two things are not necessary to produce lectal contamination. First, speakers do not need to be consciously aware of any lexical biases in their use of morphosyntactic constructions. In fact, this is unlikely anyway, since the lexical biases would expectedly be subtle and probabilistic in nature. Second, it is not required that language users know, even subconsciously, why they produce such lexical biases. The proposed explanation of lectal contamination does not entail that the Sheffielder in the example above has cognitively registered that *backpack* is a typically American word, or that the *of*-variant is more often used in American English. Instead, they might simply be reproducing the lexical biases in the language use around them, without noticing the source of the bias. This is not to say that speakers do not retain any social information about the use of words or morphosyntactic constructions – there is in fact convincing evidence that they do (Hay, 2018; Hay et al., 2019) – merely that it is not a necessary precondition for lectal contamination. In other words, the present explanation of lectal contamination does not entail that English language users prefer to use the American *of*-variant for a typically American word if they have never heard that word in the *of*-variant before – although it does not preclude such an effect either.

The present study will test the effect of lectal contamination using the Dutch partitive genitive construction as a case study, as in Examples (8) and (9). There are several formal analyses of the construction available, but the differences between these are not strictly at issue here. The most straightforward analysis that is presented in most reference grammars, views it as a combination of an indefinite pronoun, e.g. *niets* “nothing” in Example (8) or *iets* “something” in Example (9), and a postmodifying adjectival constituent, e.g. *speciaal* “special”

in Examples (8)–(9) (Haeseryn et al., 1997: 863; van den Toorn, 1977: 271). For an overview of other proposals for the formal analysis of the construction, see Pijpops and Van de Velde (2018: 103–106).

- (8) Heb je nog wat gekocht? *Niets speciaal*. (CGN file: fn008212)  
 “Have you bought anything? Nothing special.”
- (9) Maar was er *iets speciaals* of belde je zomaar? (CGN file: fn008371)  
 “But was there something special or did you just call for no reason?”

What is at issue in the present paper is the morphological variation that the construction exhibits. The adjective can appear either with or without *-s* ending, without a difference in meaning or function. Instead, the use of the *-s* ending is determined by the following factors, listed in descending order of importance: (i) more *-s* omission among a specific set of adjectives, viz. the color adjectives and the so-called ‘assessment adjectives’ *verkeerd* “wrong”, *goed* “good”, *beter* “better” and *fout* “incorrect”, which are affected by constructional contamination (see Hilpert & Flach, forthcoming; Pijpops et al., 2018; Pijpops & Van de Velde, 2016); (ii) overall more *-s* omission in the Belgian variety compared to the Netherlandic variety; (iii) more *-s* omission in informal registers; (iv) more *-s* omission among the pronouns *iets* “something” and *niets* “nothing”, although this effect appears to be restricted to the Belgian variety; and (v) more *-s* omission among partitive genitive phrases of low frequency (Pijpops & Van de Velde, 2018: 114–116). The relative importance of these last three predictors may shift somewhat and the influence of frequency seems fickle (see Pijpops & Van de Velde, 2015: 361–362, 2016: 567).

The present paper is most interested in the second factor, i.e. the distinction between the Belgian and Netherlandic variety. The influence of this factor means that the first precondition for lectal contamination, listed in Section 2, is satisfied. The second precondition is also fulfilled for the Belgian and Netherlandic Dutch varieties, as is evident from Daems et al., (2015); Geeraerts et al. (1999) and Ruetten (2012): the two lects show clear probabilistic differences in lexical usage. As for the third precondition, there is indeed language contact between Dutch speakers from the Netherlands and Belgium, given the intense economic and demographic ties between both countries (Centraal Bureau voor de Statistiek, 2020; van Agtmaal-Wobma et al., 2007). Regarding the fourth precondition, there are indications from the study of constructional contamination that language users occasionally store language chunks containing partitive genitives in memory (Pijpops & Van de Velde, 2016). In addition, there is evidence from other case studies that language users can store such ready-mades and regularly do, as mentioned above. Therefore, I expect lectal contamination to be at play for the Dutch partitive genitive, and make the following prediction. Partitive genitive phrases that are more often used in the Belgian variety will more often appear without the *-s* ending

than those that are more often used in the Netherlandic variety, even when only looking at strictly Belgian data or at strictly Netherlandic data.

### 3. Corpus study

This section presents the corpus study. Section 3.1 introduces the corpus data, while Section 3.2 describes the analyses.

#### 3.1 Data

In order to test the prediction formulated at the end of the previous section, a corpus is needed that rigidly distinguishes between Belgian and Netherlandic Dutch. The reason is that, if there is some data from Netherlandic speakers in the Belgian dataset, there is a good chance that the prediction is confirmed even if lectal contamination is not at play: Netherlandic speakers are more likely to produce the variant with *-s* ending, and they would also be more likely to use partitive genitive phrases that are typical of Netherlandic Dutch. Hence, having such speakers in the Belgian dataset could cause typically Netherlandic partitive genitive phrases to ostensibly appear more often in the typically Netherlandic variant with *-s* ending.

I have hence opted to use the Corpus of Spoken Dutch and the Sonar corpus, two corpora that aim to achieve a representative crosscut of respectively spoken and written Standard Dutch in Belgium and the Netherlands (Oostdijk et al., 2013b, 2002). The Corpus of Spoken Dutch (CGN, Corpus Gesproken Nederlands) supplies precise background information regarding the country of origin for all of its language users, while Sonar does so for some of its material. Only material with available background information was used.

To guarantee comparability with previous research, all instances in which one of the following pronouns preceded one of the following adjectives were extracted from the corpora. These lists are taken from Pijpops and Van de Velde (2018: 107). The query made use of the lemmatization of the corpora, but not of their syntactic parses, since the quality of these parses is not guaranteed for informal material (Oostdijk et al., 2013a: 49–50).

- i. Pronouns: *iets* “something”, *niets* “nothing”, *wat* “something”, *veel* “a lot”, *weinig* “few”, *zoveel* “so much”
- ii. Adjectives: *aardig* “nice”, *apart* “apart”, *belangrijk* “important”, *beter* “better”, *bijzonder* “particular”, *blauw* “blue”, *concreet* “concrete”, *deftig* “decent”, *dergelijk* “similar”, *erg* “awful”, *geel* “yellow”, *gek* “crazy”, *goed* “good”, *groen*



“green”, *interessant* “interesting”, *klein* “small”, *lekker* “tasty”, *leuk* “fun”, *mooi* “beautiful”, *nieuw* “new”, *nuttig* “useful”, *oranje* “orange”, *positief* “positive”, *purper* “purple”, *raar* “weird”, *rood* “red”, *spannend* “exciting”, *speciaal* “special”, *verkeerd* “wrong”, *verschrikkelijk* “horrible”, *vreemd* “weird”, *warm* “warm”, *wit* “white”, *zinnig* “sensible”, *zwart* “black”

Note that these pronouns and adjectives were not selected because they are markedly Belgian or Netherlandic. This constitutes a conservative choice in research design. While I would expect the effect of lectal contamination to be most clearly visible for words that are outspokenly Belgian or Netherlandic, the effect should still be present, albeit less clearly, for words that only exhibit minor lectal differences.

The instances were then manually checked to exclude any false positives, as in Pijpops and Van de Velde (2018: 108–110). 54 instances that did involve a partitive genitive, but whose adjectival constituent consisted of more than one word, such as Examples (10) and (11), were also removed. The reason is that the number of words of the adjectival constituent may form a confounding factor, for which it is hard to control, given the low number of these instances (Pijpops & Van de Velde, 2018: 122).

- (10) is voor de ouders *iets erg belangrijks*.  
 (Sonar id: WR-P-P-B-000000214.p.175.s.1)  
 “For the parents, that is *something very important*.”
- (11) Ik heb fotograferen altijd beschouwd als *iets lekker stouts*.  
 “I have always considered making photographs *something delightfully naughty*.”  
 (Sonar id: WR-U-E-A-0000104133.text.1.event.3910.s.1)

This left me with a total of 9984 instances, of which 8438 stem from Netherlandic language users, and 1546 from Belgian language users. The reason why there are a lot more Netherlandic than Belgian observations is that Sonar supplies much more Netherlandic data for which the writer is known than Belgian data. This imbalance is actually fortuitously convenient, as the distribution between both variants is much more uneven in the Netherlands than in Belgium. The final dataset contains 908 instances with *-s* vs. 638 without *-s* from Belgian language users, and 7784 with *-s* vs. 654 without *-s* from Netherlandic language users.

### 3.2 Analyses

All statistical analyses were executed in *R* (R Core Team, 2014), using the packages *lme4* (Bates et al., 2013), *effects* (Fox et al., 2016), *vcd* (Meyer et al., 2020) and *Hmisc* (Harrell, 2017). Based on the entire final dataset, I calculated for each

partitive genitive phrase, i.e. for each unique combination of a pronoun and an adjective, such as *veel leuk(s)* “a lot of fun things”, *iets bijzonder(s)* “something special” or *niets speciaal(s)* “nothing special”, how typically Belgian or Netherlandic it is, regardless of whether it is used with or without the *-s*. This was done as in Equation 1. I took the logarithm with base 10 of the ratio between the odds of the phrase in the Belgian dataset and the odds of the phrase in the Netherlandic dataset. In order to avoid division by zero, I added 1 to the count of the phrase in the Netherlandic dataset. For balance, I also added 1 to the count of the phrase in the Belgian dataset. The higher this log odds ratio, the more “typically Belgian” a phrase is, the lower the log odds ratio, the more “typically Netherlandic” a phrase is. I hence called this measure `BELGIANNES PHRASE` and predicted it to be positively correlated with the probability of *-s* omission.

Equation 1. Calculation of `BELGIANNES PHRASE`

$$\text{BELGIANNES PHRASE}_{\text{PHRASE}} = \log \left( \frac{(\text{count of phr. } x \text{ in the Belg. dataset} + 1) / (\text{total size of the Belg. dataset} - (\text{count of phr. } x \text{ in the Belg. dataset} + 1))}{(\text{count of phr. } x \text{ in the Neth. dataset} + 1) / (\text{total size of the Neth. dataset} - (\text{count of phr. } x \text{ in the Neth. dataset} + 1))} \right)$$

To test this prediction, two logistic mixed regression models were built: one on the Belgian data and one on the Netherlandic data. Both had the use of the *-s* as their response variable, with *-s* omission being the success level. The predictors included `BELGIANNES PHRASE` as a fixed effect, and `PHRASE` as a random effect. `PHRASE` is a variable with a separate level for each partitive genitive phrase. This random effect had to be added to the model because `BELGIANNES PHRASE` has an identical value for all observations of the same phrase (Speelman et al., 2018: 2).

From a purely statistical point of view, it would make sense to build a single regression model on the entire dataset containing both the Belgian and Netherlandic data, especially since the effect of `BELGIANNES PHRASE` is predicted to be the same for the Belgian data and the Netherlandic data. That model could then simply control for the lectal distinction by integrating `COUNTRY` as a fixed effect. However, I prefer to build separate regression models for the lects for conceptual reasons, to drive home the point that lectal contamination entails the existence of lexical biases *within* individual lects.

The following variables were also added as fixed effects to the regression models (Pijpops & Van de Velde, 2018: 121). `TYPE ADJECTIVE` distinguishes between the color adjectives and the assessment adjectives, which are affected by constructional contamination, and all other adjectives. `REGISTER` distinguishes between a formal and an informal register. For the CGN, I made use of the division proposed in Plevoets (2008: 80). This involves viewing the components a, c, and d, viz. the spontaneous conversations and telephone conversations, as informal, and the other components as formal. For Sonar, the material from the components `WRPEA`, `WRPEL`, `WRUEA` and `WRUED`, viz. the discussion lists, tweets, chat

logs, and text messages, was labeled as informal, while the material from the components WREI, WRPPB, WPPPH, WRPPK and WRUEE, viz. the websites, books, periodicals & magazines, reports, and written assignments, was labeled as formal. None of the other components of the corpus contained material for which the country of the author was known. Finally, the variables CORPUS distinguishes between the material from the CGN and the Sonar corpus.

- TYPE ADJECTIVE: *assessment, color, other*
- REGISTER: *formal, informal*
- PRONOUN: *iets* “something”, *niets* “nothing”, *veel* “a lot”, *wat* “something”, *weinig* “few”, *zoveel* “so much”
- FREQUENCY: log-transformed frequency of the partitive genitive phrase in the dataset
- CORPUS: *CGN, Sonar*

All categorical predictors were implemented through dummy coding. The models were then fitted to the Netherlandic and Belgian data. All variance inflation factors were well below 5 (Levshina, 2015:160), and a binned residual analysis for each model did not reveal any anomalous patterns (Gelman & Hill, 2007:97–98; Sonderegger et al., 2018: Section 5.4.1). The specifications of the models can be found respectively in Table 1 and Table 2, and the effect plots of BELGIANNES PHRASE can be found in Figure 2. The model based on the Netherlandic data has a C-index of 0.922, which indicates outstanding predictive quality according to Hosmer and Lemeshow (2000:162), while the model based on the Belgian data has a C-index of 0.779, indicating acceptable predictive quality. Tables 1–2 also report the percentage of occurrences for which the variant was correctly predicted by the model, as well as the baseline. Still, C-indices are generally preferred as an indication of predictive quality since they can be compared across different baselines (Speelman, 2014: 513–515).

Both models show a significant positive effect of BELGIANNES PHRASE on -s omission, albeit that the effect is only barely significant in the Belgian model. This confirms the prediction of lectal contamination. The other predictors indicate the same effects as found in previous research, and CORPUS shows that in the Belgian model, -s omission is more often used in the spoken material of the CGN than in the written material from Sonar. This may be interpreted as the variant without -s functioning as the colloquial variant in Belgian Dutch (Pijpops & Van de Velde, 2018: 121).

**Table 1.** Specifications of the regression model fitted on the Netherlandic data

AIC:	2793.3	Number of observations with -s:	7784
C-index:	0.922	Number of observations without -s (success level):	654
Correctly predicted:	0.95	Baseline:	0.92

Fixed effects	Level	Estimate	Standard error	Z-value	P-value
	intercept	-2.80	1.04	-2.70	0.0070
BELGIANNES PHRASE		1.57	0.49	3.20	0.0014
TYPE ADJECTIVE	assessment	Reference level			
	color	2.01	0.67	3.01	0.0026
	other	-1.97	0.47	-4.16	<0.0001
REGISTER	formal	Reference level			
	informal	0.69	0.12	5.95	<0.0001
PRONOUN	<i>iets</i> “something”	Reference level			
	<i>niets</i> “nothing”	0.94	0.57	1.66	0.0961
	<i>veel</i> “a lot”	2.47	0.57	4.35	<0.0001
	<i>wat</i> “something”	1.37	0.55	2.48	0.0132
	<i>weinig</i> “few”	2.21	0.71	3.10	0.0019
	<i>zoveel</i> “so much”	1.05	0.83	1.26	0.2086
FREQUENCY		-0.08	0.15	-0.54	0.5924
CORPUS	Sonar	Reference level			
	CGN	-0.26	0.25	-1.04	0.3000

Random effect	Number of levels	Variance	Standard deviation
PHRASE	165	2.24	1.50

#### 4. Agent-based simulation

The prediction based on lectal contamination has been confirmed in the previous section. However, the corpus study does of course suffer from the unavoidable drawbacks of corpus research. First, it is in principle possible that the correlations shown in Figure 2 are caused by some unknown factor that the regression models failed to control for. Second, the quality of the analyses is dependent on the quality of the corpus. Any mistakes made during its compilation could affect the present results. Meanwhile, apart from these corpus findings, the argument for lectal contamination is currently only based on verbal reasoning, which is ill-suited to fully understand emergent effects.

**Table 2.** Specifications of the regression model fitted on the Belgian data

AIC:	1801.9	Number of observations with -s:	908
C-index:	0.779	Number of observations without -s (success level):	638
Correctly predicted:	0.72	Baseline:	0.59

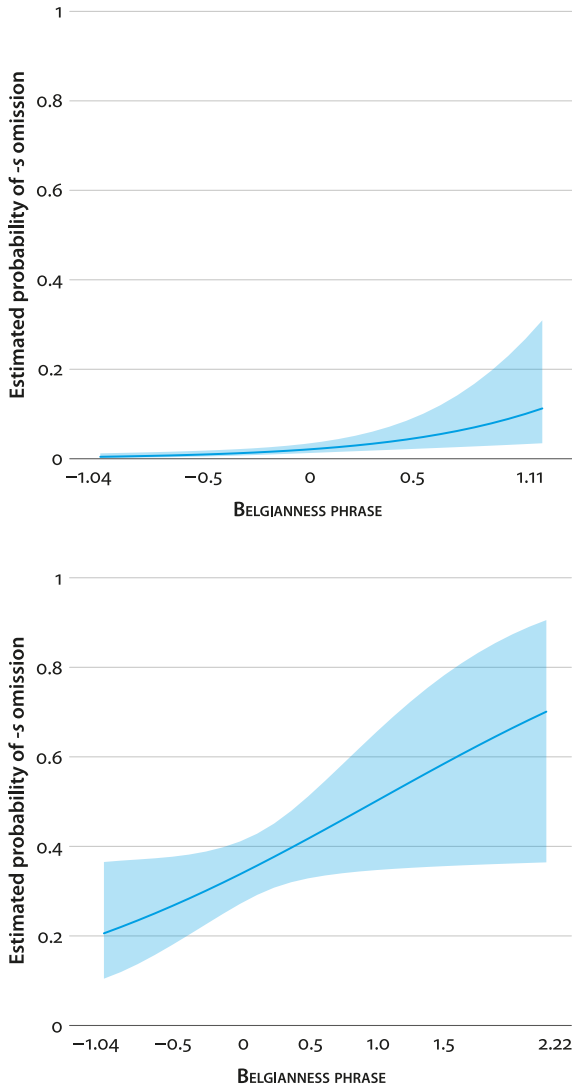
  

Fixed effects	Level	Estimate	Standard error	Z-value	P-value
	intercept	1.37	0.92	1.49	0.1367
BELGIANNES PHRASE		0.67	0.33	2.04	0.0414
TYPE ADJECTIVE	assessment	Reference level			
	color	1.12	0.66	1.69	0.0912
	other	-1.93	0.35	-5.57	<0.0001
REGISTER	formal	Reference level			
	informal	1.32	0.33	4.05	0.0001
PRONOUN	<i>iets</i> “something”	Reference level			
	<i>niets</i> “nothing”	-0.96	0.43	-2.26	0.0239
	<i>veel</i> “a lot”	-0.82	0.43	-1.89	0.0585
	<i>wat</i> “something”	-0.79	0.46	-1.71	0.0881
	<i>weinig</i> “few”	-1.41	0.76	-1.85	0.0639
	<i>zoveel</i> “so much”	-1.35	0.80	-1.69	0.0916
FREQUENCY		-0.31	0.14	-2.29	0.0223
CORPUS	Sonar	Reference level			
	CGN	0.38	0.20	1.97	0.0494

Random effect	Number of levels	Variance	Standard deviation
PHRASE	120	0.69	0.83

Still, the four theoretical preconditions that are listed in Section 2 are well-established in the literature. If it can be shown that the effect of lectal contamination has to emerge under these preconditions, then the argument for lectal contamination becomes a lot stronger. This is exactly what the present section will try to do by using agent-based modelling. I first introduce the conceptual design of the simulation and discuss how the simulation will be evaluated, and then turn to its practical implementation. Finally, the results are presented.



**Figure 2.** Effect plots of BELGIANNES PHRASE (top: The Netherlandic model; bottom: The Belgian model)

#### 4.1 Design and evaluation

The simulation should consist of the following elementary building blocks. These building blocks are directly derived from the four theoretical preconditions for lectal contamination presented in Section 2.

- i. 2 communities of agents
- ii. 2 morphosyntactic variants, with differing initial, not-hardcoded preferences in each community
- iii. 2 lexical items, with differing initial, not-hardcoded preferences in each community
- iv. Occasional language contact between the communities
- v. Cognitive storage of ready-made language forms
- vi. Experience affecting cognitive entrenchment
- vii. Cognitive entrenchment affecting usage
- viii. No initial lectal contamination

While agent-based modelling presents an effective way of clarifying emergent effects by way of simulating them, it is certainly not my goal to build an ultimately realistic simulation of human communication. Instead, I merely mean to determine the minimal conditions under which lectal contamination has to take place. This means that I should strive to build the simplest simulation possible. For any implementation choice that is not dictated by these eight building blocks, I have therefore chosen the simplest option, in accordance with best practices in agent-based modelling (Landsbergen, 2009: 18–19; van Trijp & Steels, 2012: 9).

The following paragraphs discuss these building blocks in the order presented above. The simulation consists of two communities of agents that communicate with each another by means of two morphosyntactic variants and two lexical items. An agent of the first community starts with an initial relative preference for the first morphosyntactic variant and the first lexical item, while an agent of the second community starts with an initial relative preference for the second morphosyntactic variant and the second lexical item. These preferences are not hard-coded. That is, they may alter during a run of the simulation as a result of the interactions in which the agents partake. An agent of the first community may occasionally interact with an agent of the second community and vice versa, but most interactions take place between agents of the same community.

The agents store ready-made forms in their memory, i.e. combinations of a lexical item and a morphosyntactic variant. The more often an agent hears a form, the better it is entrenched in the agent's memory. The better a form is entrenched in an agent's memory, the more likely the agent is to use it in a future interaction. Finally, there should not be any lectal contamination present at the start of the simulation. The morphosyntactic and lexical preferences of the communities should initially be exactly independent of each other, and no lexical biases should be present at the start.

If it can be shown that lectal contamination emerges under these conditions, the simulation will be evaluated as being successful. That is, if under these condi-

tions, the first lexical item consistently develops a relative preference for the first morphosyntactic variant, while the second lexical item consistently develops a relative preference for the second morphosyntactic variant among the agents of both communities, then the simulation will have shown that lectal contamination has to emerge under the four theoretical preconditions listed in Section 2.

## 4.2 Implementation

The simulation is implemented in *Python* (van Rossum & Drake, 2009) such that the conceptual building blocks in Subsection 4.1 are respected. For reasons of clarity, I call that the first morphosyntactic variant *with -s*, and the second morphosyntactic variant *without -s*, while referring to the first lexical item as *iets bijzonder(s)* and to the second lexical item as *iets speciaal(s)*. These are only names, however, that do not affect the functioning of the simulation in any way: I could have just as well named the variants Variant A and Variant B, or *with of* versus *without of*, and the lexical items Lexeme A and Lexeme B, or *backpack* versus *rucksack*.

An agent retains in its memory an inventory of ready-made forms, with a maximum of four, corresponding to each possible combination of a morphosyntactic variant and a lexical item, i.e. *iets bijzonders*, *iets bijzonder*, *iets speciaals*, *iets speciaal*. For each form in memory, a count is also retained, corresponding to the number of times the agent has heard the form. The higher the count, the more strongly the form is entrenched in the agent's memory.

The agents are divided into two communities, viz. Community A and Community B. At the start of the simulation, before any interactions have taken place, the sum of all counts in the memory of each agent equals 100. For an agent of Community A,  $m_A$  of these counts are allotted to the first morphosyntactic variant, with  $100 - m_A$  being allotted to the second variant, and  $l_A$  of these counts are allotted to the first lexical item, with  $100 - l_A$  being allotted to the second item. For an agent of Community B,  $m_B$  of these counts are allotted to the first morphosyntactic variant, with  $100 - m_B$  being allotted to the second variant, and  $l_B$  of these counts are allotted to the first lexical item, with  $100 - l_B$  being allotted to the second item.

At the start of the simulation, the agents of Community A should have a preference for one of the morphosyntactic variants and one of the lexical items, relative to the agents of Community B. In practice, this means that the following must always hold:  $m_A \neq m_B$  and  $l_A \neq l_B$ . For instance, in the simulational runs presented below, I have set these parameters as  $m_A=100$ ,  $l_A=80$ ,  $m_B=60$ ,  $l_B=20$ . In effect, this means that the agents of Community A start with a memory like this:  $\{ \textit{iets}$



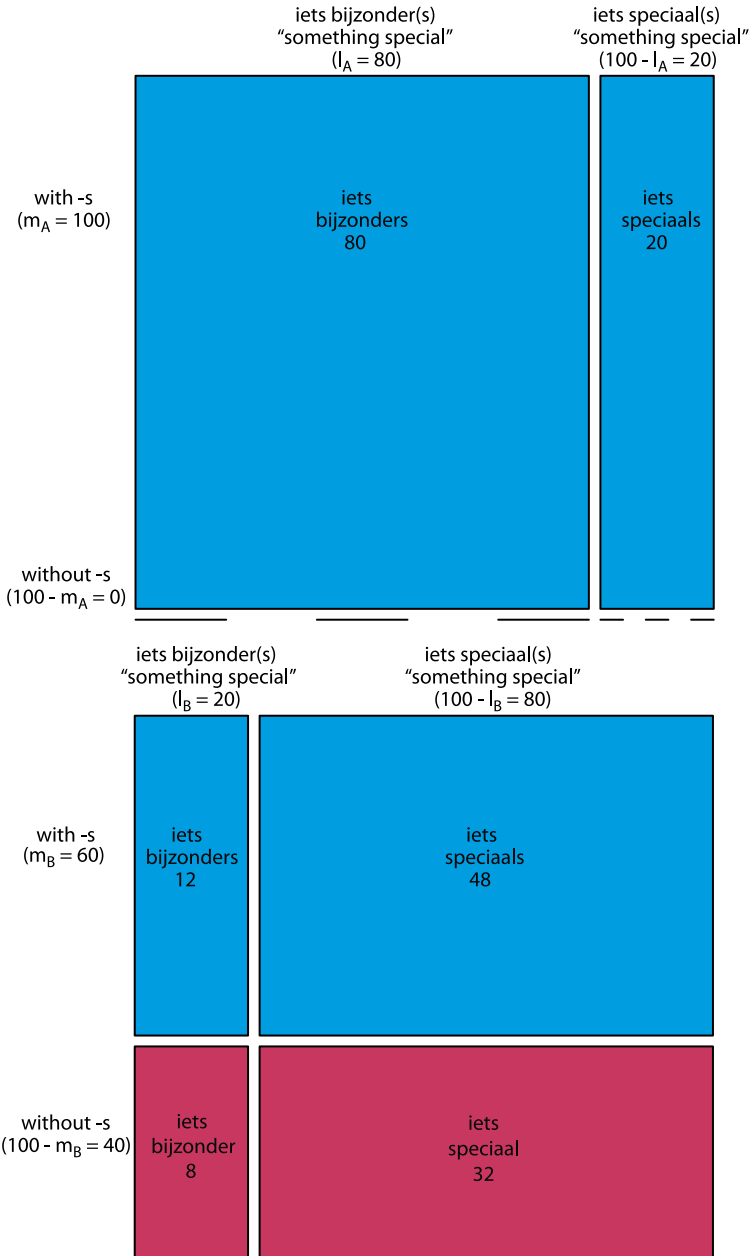
*bijzonders: 80, iets speciaals: 20* }, while the agents of Community B start with a memory like this *{ iets bijzonders: 12, iets bijzonder: 8, iets speciaals: 48, iets speciaal: 32 }*. Note that the initial morphosyntactic and lexical preferences are exactly independent of each other, as can be seen in the mosaic plots in Figure 3. In other words, there is no lectal contamination and indeed no lexical bias whatsoever present at the start of the simulation. I have so far implemented building blocks (i), (ii), (iii), (v) and (viii).

At each “point in time”,  $n$  interactions between two agents take place, with  $n$  equal to the total number of agents in the simulation. For  $a \cdot n$  of these interactions, the speaker agent will be selected from Community A, and for  $(1 - a) n$  of these interactions, the speaker agent will be selected from Community B, with  $a$  equal to the percentage of the total population that is part of Community A. In the simulational runs presented below,  $n$  is always set to 100, with  $a$  set to 0.5. For each interaction between two agents, there is a probability  $h$  that the hearer agent is selected from a different community than the speaker agent. As long as this parameter  $h$  is set to some value between 0 and 0.5, this implements building block (iv).

Finally, Figure 4 represents an example of an interaction between two agents. The speaker agent selects a form  $x$  from its memory with probability  $p_x$ , based on the count  $c_x$  of that form in its memory, as in Equation 2. When the hearer agent has never heard that form before, it adds it to its memory with count 1. If the hearer agent has heard the form before, it simply adds 1 to its count. Only the hearer updates the counts in its memory, not the speaker (see De Vylder, 2007; van Trijp & Steels, 2012). This implements building blocks (vi) and (vii). There are many other, arguably more realistic, ways of implementing these building blocks. Overviews of some of the formulas that have been used to select an utterance from a speaker’s memory and update a hearer’s memory can be found in Pijpops and Beuls (2015: 13–19) and Wellens (2012: 33–136). The choice between these formulas is not dictated by any of the building blocks: all of them would be in accordance with building blocks (vi) and (vii). I therefore went with the implementation that I deemed simplest (see Landsbergen, 2009: 18–19).

Equation 2. Probability  $p_x$  of a speaker agent selecting form  $x$  with count  $c_x$  in its memory

$$p_x = \frac{c_x}{\sum_{i=1}^4 c_i}$$



**Figure 3.** Distributions of the counts in the initial memories of the agents (top: For agents of Community A; bottom: For agents of Community B)

Memory		Probabilities		Utterance	Memory	
form $x$	count $c_x$	form $x$	$p_x$		form $x$	count $c_x$
<i>iets bijzonders</i>	100	<i>iets bijzonders</i>	0.5	→ “iets speciaals” →	<i>iets bijzonders</i>	83
<i>iets bijzonder</i>	30	<i>iets bijzonder</i>	0.15		<i>iets bijzonder</i>	61
<i>iets speciaals</i>	60	<i>iets speciaals</i>	0.3		<i>iets speciaals</i>	35 + 1
<i>iets speciaal</i>	10	<i>iets speciaal</i>	0.05		<i>iets speciaal</i>	11

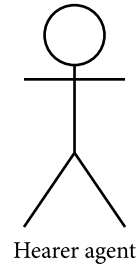
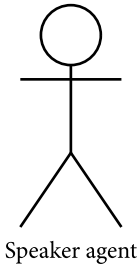


Figure 4. Example of an interaction between a speaker agent and a hearer agent

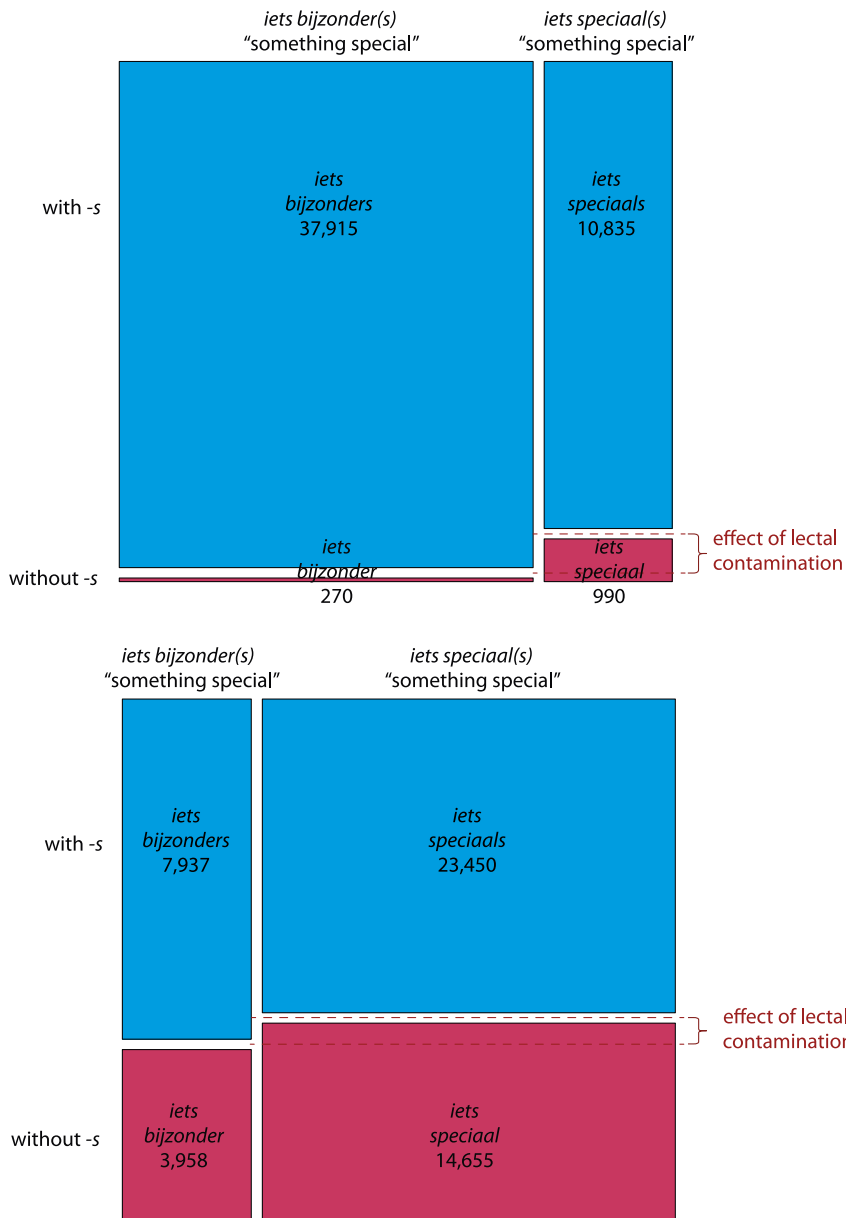
### 4.3 Results

The simulation was run for 100000 points in time, which corresponds to 10 million interactions, with the following parameters.

- Initial morphosyntactic preferences:  $m_A=100$ ,  $m_B=60$
- Initial lexical preferences:  $l_A=80$ ,  $l_B=20$
- Total population size:  $n=100$
- Percentage of the total population that is part of Community A:  $a=0.5$
- Language contact:  $h=0.01$

If lectal contamination emerges, a lexical bias is expected to emerge such that the variant with *-s* becomes used more often among the occurrences of the lexical item *iets bijzonder(s)*, while the variant without *-s* becomes used more often among the occurrences of the lexical item *iets speciaal(s)*, both among the utterances of the agents of Community A and the agents of Community B. Figure 5 shows the number of times each form was produced by agents of Community A and Community B during the final 1000 points in time of the simulational run. The results indeed show the predicted effect (Community A:  $p < 0.0001$ , Cramer's  $V=0.21$ , Community B:  $p < 0.0001$ , Cramer's  $V=0.05$ ).

Still, since the simulation makes extensive use of probabilities, the results of a simulational run will be slightly different each time one is executed, even when the exact same parameter settings are chosen. In order to investigate whether the results presented in Figure 5 prove consistent, a batch of 100 simulational runs with the exact same parameter settings was executed. To track the degree of lectal



**Figure 5.** Distributions of the number of times each form was produced during the final 1000 points in time of a single simulational run (top: By agents of Community A; bottom: By agents of Community B)

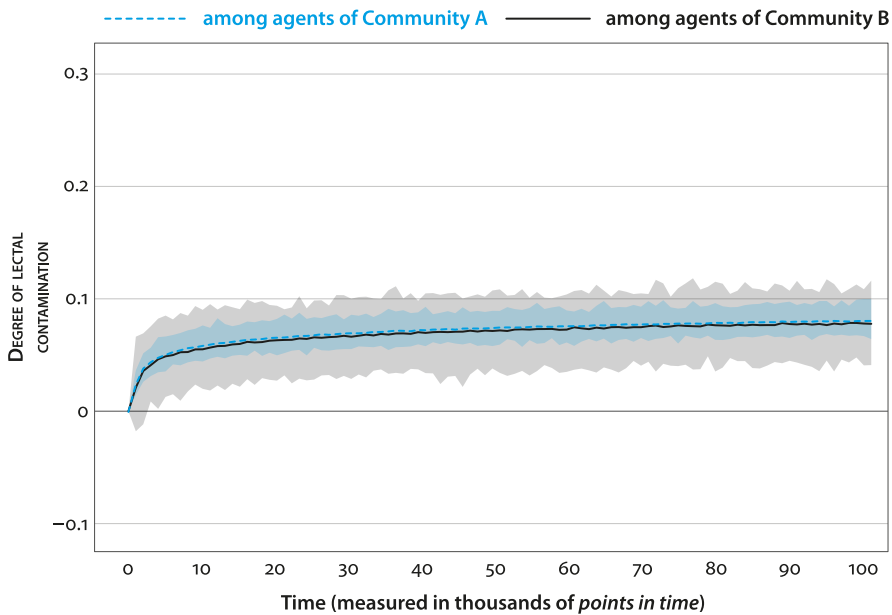
contamination present in the simulation, the measure DEGREE OF LECTAL CONTAMINATION was recalculated every 1000 points in time in each simulational run, as

in Equation 3. The value of this measure corresponds to the size of the red bracket in Figure 5.

Figure 6 plots DEGREE OF LECTAL CONTAMINATION through time, averaged over 100 simulational runs, among the utterances produced by the agents of Community A and Community B. The error bars indicate the minimum and maximum values over those 100 runs. The results show that the effect of lectal contamination indeed emerges consistently.

Equation 3. Calculation of the degree of lectal contamination present in the simulation. This measure is recalculated every 1000 points in time

$$\text{DEGREE OF LECTAL CONTAMINATION} = \frac{\text{usages of iets bijzonders}}{\text{usages of iets bijzonders} + \text{iets bijzonder}} - \frac{\text{usages of iets speciaals}}{\text{usages of iets speciaals} + \text{iets speciaal}}$$



**Figure 6.** Development of lectal contamination through time in 100 simulational runs, among the utterances of agents of Community A and Community B

The simulation can now be used to test whether each of the preconditions listed in Section 2 are indeed necessary for lectal contamination. Figure 7 shows the development of lectal contamination among the agents of Community A when each of these preconditions is removed. The first precondition can be removed by setting the parameters  $m_A = m_B$ . This entails that there is no initial difference between the lects regarding their relative preference for one of the morphosyntactic variants. Likewise, the second precondition can be removed from the simulation by setting

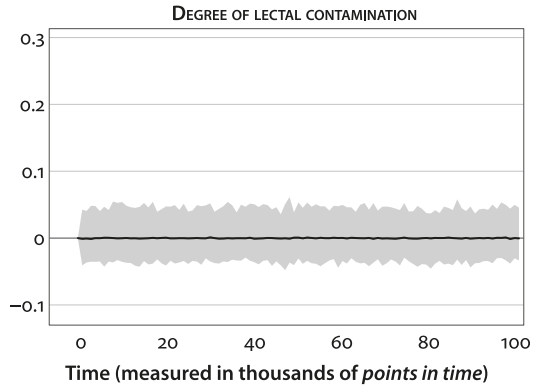
the parameters  $l_A = l_B$ . This entails that there is no initial difference between the lects regarding their relative preference for one of the lexical items. The third precondition can be removed by setting the parameter  $h=0$ . This entails that an agent of Community A never interacts with an agent of Community B and vice versa, effectively banning language contact from the simulation.

Finally, removing the fourth precondition requires a more drastic change to the simulation, as it constitutes a pivotal theoretical issue pertaining to how grammar is cognitively structured. Instead of the agents storing ready-made forms in memory, they are now equipped with a separate inventory for the morphosyntactic variants and for the lexical items. When a speaker agent then needs to produce a form, it first calculates the probabilities of each variant, chooses a variant according to these probabilities, and then does the same for the lexical items. Finally, it produces the language form that corresponds to the chosen variant and the chosen lexical item. Similarly, a hearer agent dissects the heard language form and adds 1 to the count of the morphosyntactic variant and 1 to the count of the lexical item that it has heard. As such, no ready-made forms are stored in the memory of the agents. Figure 7 shows that if any one of these preconditions is removed, lectal contamination does not emerge in the simulation.

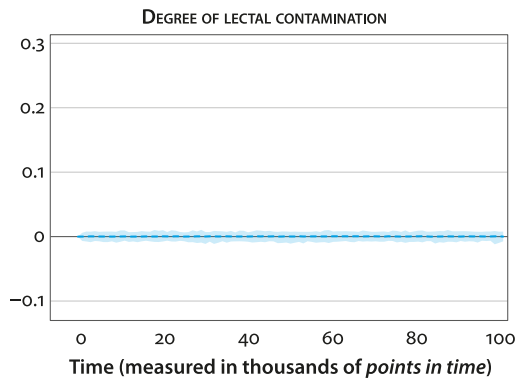
Finally, each of the parameter settings is systematically varied. Figure 8 shows the development of lectal contamination in cases where there is more or less language contact between the communities, i.e. it shows DEGREE OF LECTAL CONTAMINATION for several settings of the parameter  $h$  with  $h > 0$ . It is found that lectal contamination increases if there is more language contact in the simulation, but consistently emerges as long as there is some minor degree of language contact present, i.e. as long as  $h > 0$ . Additional figures in the appendix show the results of varying the other parameters on the DEGREE OF LECTAL CONTAMINATION during the final 1000 points in time of simulational runs of 100000 points in time. It is found that, although the size of the effect evidently varies as a function of particular parameters, lectal contamination as such does emerge consistently as long as all four preconditions are fulfilled.

## 5. Discussion and conclusions

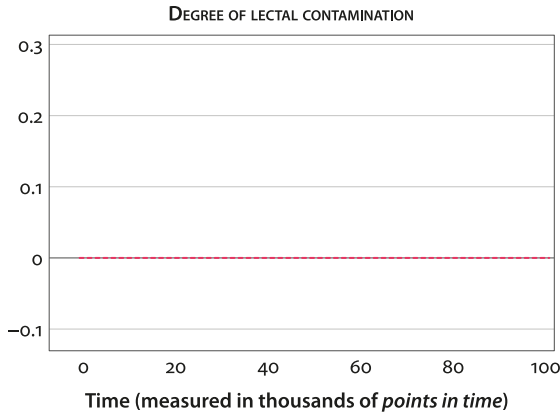
The first goal of this paper was to investigate the effect of lectal contamination. The paper has presented both observational and simulational evidence for this effect. It has shown that (i) the effect appears to be present in the real world for the Dutch partitive genitive construction; and (ii) the effect consistently emerges in an agent-based simulation under four key theoretical preconditions. This indicates that if these four preconditions are met in reality, lectal contamination



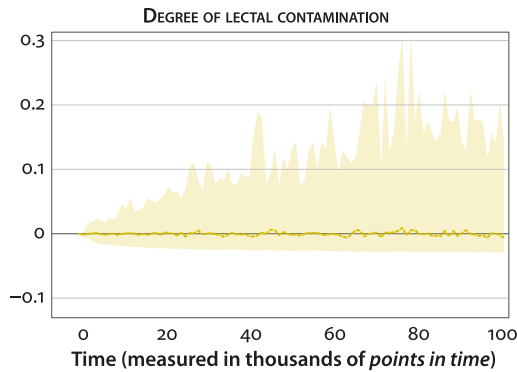
- a. Precondition (i) removed, no difference in initial morphosyntactic preference:  $m_A = m_B = 50$



- b. Precondition (ii) removed, no difference in initial lexical preference:  $l_A = l_B = 50$



c. Precondition (iii) removed, no language contact:  $h = 0$



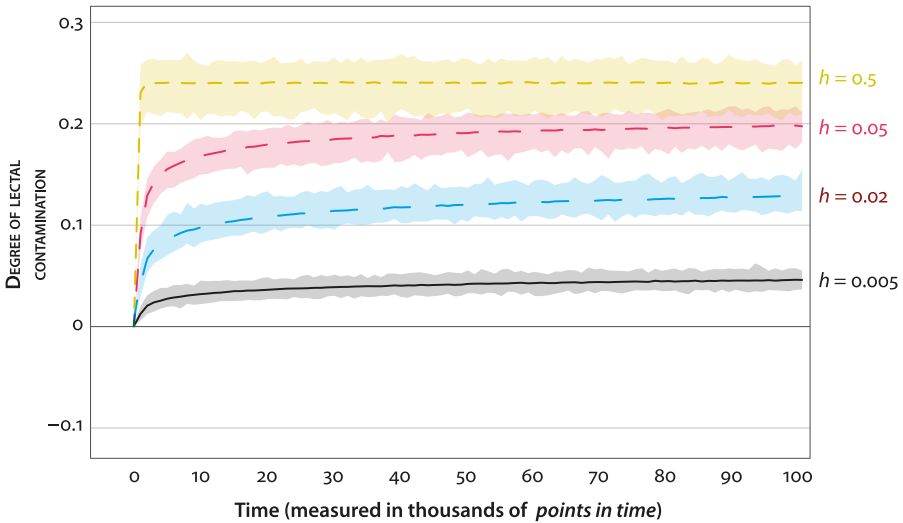
d. Precondition (iv) removed, no storage of ready-made forms

**Figure 7.** Development of lectal contamination through time in 100 simulational runs, among the utterances of agents of Community A, when each of the preconditions is removed

should develop. Still, reality is of course infinitely more complex than the simulation presented in this paper. It is in principle possible that, while the four preconditions should lead to the emergence of lectal contamination, some other element or mechanism that is not implemented in the simulation but that does exist in the real world is blocking the rise of lectal contamination. I am currently unaware of any element that may do so, however.

Lectal contamination is important for four reasons. First, there is a potentially high number of affected cases of morphosyntactic variation. The four preconditions are very general in nature, and conceivably cover various morphosyntactic





**Figure 8.** Development of lectal contamination through time in 100 simulational runs, among the agents of Community A, for different settings of the language contact parameter  $h$

alternations, some of which are already mentioned in the introduction of this paper. More examples can be found in the overviews of Algeo (2006), Haeseryn (2013) or Dürscheid et al. (2018).

Second, lectal contamination provides a mechanism that can create lexical biases in morphosyntactic variation, and the importance of such biases is increasingly underlined in recent work. These biases may form the basis from which meaning differences between morphosyntactic variants develop (Coleman, 2009; Perek & Goldberg, 2015; Pijpops, 2019: 63–86). Moreover, they testify of the constructional links between schematic constructions and their slot fillers (Diessel, 2019: 113–195).

Third, lectal contamination demonstrates how a language-external cause, i.e. a mere difference in the base distribution of variants in two or more lects, can have language-internal effects; it can create lect-internal lexical biases. This shows how language contact does not only have directly visible effects on the lexicon, but also affects morphosyntax in subtle, probabilistic ways that typically remain undetected to the naked eye. As such, lectal contamination reinforces the argument of Höder (2014, 2018), that language contact should not be viewed as a purely extralinguistic phenomenon and that multilingualism or *multilectalism* is to be integrated into the design of the language system itself.

Fourth, lectal contamination provides additional evidence for the claim that language users cognitively store and process ready-made language forms in mem-

ory (Arnon & Snider, 2010; Dąbrowska, 2014; Tremblay et al., 2011). As can be seen by comparing Figure 6 to Figure 7d, it forms a crucial assumption to explain lectal contamination.

So far, the only observational case study of lectal contamination, i.e. the Dutch partitive genitive, is morphological in nature. However, if the four preconditions are valid for a case of syntactic variation, the effect should also emerge there. In this regard, the most questionable precondition of these four is probably the final one, concerning the storage of ready-made forms. Still, there are strong indications that language users are at least capable of occasionally storing syntactic ready-mades in memory (Arnon & Snider, 2010; Tremblay & Baayen, 2010).

Future research on lectal contamination can take various forms. For example, such research could investigate additional observational case studies, or could study how lectal contamination behaves when there is (strongly) unbalanced language contact, such that speakers of one lect much more often hear language use from the other lect than vice versa. For example, perhaps Britons watch American TV shows more often than Americans watch British ones. This could result in Britons being more exposed to American English than Americans are to British English. Another possibility is to look at the lectal awareness of language users. Language users do not *need* to be aware of the lectal biases in the distribution of morphosyntactic variants or lexical items in order for lectal contamination to take place, as explained in Section 2. Still, lectal contamination does not preclude such awareness either. A possible experiment to investigate this matter, could, for instance, teach an American neologism to British speakers, without using it in the morphosyntactic construction under scrutiny. Next, it could be tested whether the participants immediately start to prefer morphosyntactic variants of American English when using that word. If lectal contamination is only driven by ready-made cognitive storage and not at all by lectal awareness, this would not be expected to be the case.

The second goal of the paper was to present agent-based modelling as a technique that can be usefully combined with observational studies such as corpus research in order to investigate language variation. While corpus research shows us the biases and tendencies present in real world language variation, agent-based modelling allows us to simulate the mechanisms that are theorized to cause these biases and tendencies (see Landsbergen et al., 2010:367–368). When both are combined, it becomes possible to directly compare observational and simulated results, as was done in the present paper. I am therefore convinced that agent-based modelling can form a useful addition to the methodological toolbox of variational linguists, as it has already proven to be for researchers in evolutionary and historical linguistics.

## Funding

Research funded by Onderzoeksraad, KU Leuven to Dirk Pijpops.

## Acknowledgments

I am greatly indebted to Freek Van de Velde, whose assistance has been instrumental in developing my ideas on lectal contamination. In addition, I would like to thank Dirk Speelman, Karlien Franco, and Gert De Sutter for interesting discussions and useful feedback, as well as Katrien Beuls for introducing me to agent-based simulation. I am also grateful to the anonymous reviewers and the editor, whose comments have greatly helped to improve this paper.

## References

- Algeo, J. (2006). *British or American English?: A Handbook of Word and Grammar Patterns*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511607240>
- Arnon, I., & Snider, N. (2010). More than words: frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67–82. <https://doi.org/10.1016/j.jml.2009.09.005>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4* (Version 1.4) [Computer software]. <http://cran.r-project.org/package=lme4>
- Beckner, C., Blythe, R., Bybee, J., Christiansen, M., Croft, W., Ellis, N., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, 59(1), 1–26. <https://doi.org/10.1111/j.1467-9922.2009.00533.x>
- Bentivoglio, P., & Sedano, M. (2011). Morphosyntactic variation in Spanish-speaking Latin America. In M. Díaz-Campos (Ed.), *The Handbook of Hispanic Sociolinguistics* (pp. 123–147). Blackwell. <https://doi.org/10.1002/9781444393446.ch8>
- Beuls, K., & Steels, L. (2013). Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLoS ONE*, 8(3), e58960. <https://doi.org/10.1371/journal.pone.0058960>
- Bloem, J., Versloot, A., & Weerman, F. (2015). An agent-based model of a historical word order change. In R. Berwick, A. Korhonen, A. Lenci, T. Poibeau, & A. Villavicencio (Eds.), *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning* (pp. 22–27). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W15-2404>
- Blythe, R., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, 88(2), 269–304. <https://doi.org/10.1353/lan.2012.0027>
- Broekhuis, H. (2013). *Syntax of Dutch: Adjectives and Adjective Phrases*. Amsterdam University Press.
- Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511750526>

- Bybee, J. (2013). Usage-based theory and exemplar representations of constructions. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 49–69). Oxford University Press.
- Centraal Bureau voor de Statistiek. (n.d.). Retrieved March 13, 2020, from <https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83926NED/barv?dl=17256&ts=1584114358740>
- Claes, J. (2015). Competing constructions: The pluralization of presentational *haber* in Dominican Spanish. *Cognitive Linguistics*, 26(1), 1–30. <https://doi.org/10.1515/cog-2014-0006>
- Colleman, T. (2009). Verb disposition in argument structure alternations: A corpus study of the dative alternation in Dutch. *Language Sciences*, 31(5), 593–611. <https://doi.org/10.1016/j.langsci.2008.01.001>
- Dąbrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics*, 25(4), 617–653. <https://doi.org/10.1515/cog-2014-0057>
- Daems, J., Heylen, K., & Geeraerts, D. (2015). Wat dragen we vandaag: een hemd met blazer of een shirt met jasje? [What do we wear today: A 'hemd' with a 'blazer' or a 'shirt' with a 'jasje'?] *Taal En Tongval*, 67(2), 307–342. <https://doi.org/10.5117/TET2015.2.DAEM>
- Davies, M. (2004). *British National Corpus (from Oxford University Press)*. Retrieved January, 2020, from <https://www.english-corpora.org/bnc/>
- Davies, M. (2008–). *The Corpus of Contemporary American English (COCA)*. Retrieved January, 2020, from <https://www.english-corpora.org/coca/>
- De Vylder, B. (2007). *The Evolution of Conventions in Multi-agent Systems* [Doctoral dissertation, Vrije Universiteit Brussel]. <https://langev.com/pdf/deVylder07evolutionOfConventionsPHD.pdf>
- Diessel, H. (2015). Usage-based construction grammar. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 296–322). De Gruyter Mouton. <https://doi.org/10.1515/9783110292022-015>
- Diessel, H. (2019). *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge University Press. <https://doi.org/10.1017/9781108671040>
- Dürscheid, C., Elspaß, S. & Ziegler, A. (Eds.). (2018). *Varietengrammatik des Standarddeutschen. Ein Online-Nachschlagewerk* [Variant grammar of Standard German. An online reference work]. [http://mediawiki.ids-mannheim.de/VarGra/index.php/Substantive\\_auf\\_ation/\\_-ung](http://mediawiki.ids-mannheim.de/VarGra/index.php/Substantive_auf_ation/_-ung)
- Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, 120(8), 2061–2079. <https://doi.org/10.1016/j.lingua.2010.02.001>
- Fox, J., Weisberg, S., Friendly, M., Hong, J., Andersen, R., Firth, D., & Taylor, S. (2016). *Effect Displays for Linear, Generalized Linear, and Other Models (Version 3.2)* [Computer software]. <https://cran.r-project.org/web/packages/effects/>
- Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen* [Convergence and divergence in Dutch vocabulary: A study into clothing and football terminology]. P. J. Meertens-Instituut.
- Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Gilbert, N. (2008). *Agent-based Models*. Sage. <https://doi.org/10.4135/9781412983259>
- Haeseryn, W. (2013). Belgian Dutch. In F. Hinskens & J. Tældeman (Eds.), *Language and Space: Dutch* (pp. 700–720). De Gruyter Mouton.

- Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., & van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst* [General Dutch Grammar]. Nijhoff.
- Harrell, F.J. (2017). *Hmisc: Harrell Miscellaneous* (Version 4.0-3) [Computer software]. <https://cran.r-project.org/package=Hmisc>
- Hay, J. (2018). Sociophonetics: The role of words, the role of context, and the role of words in context. *Topics in Cognitive Science*, 10(4), 696–706. <https://doi.org/10.1111/tops.12326>
- Hay, J., Walker, A., Sanchez, K., & Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLoS ONE*, 14(2), e0210793. <https://doi.org/10.1371/journal.pone.0210793>
- Hilpert, M., & Flach, S. (forthcoming). A case of constructional contamination in English: Modified noun phrases influence adverb placement in the passive. In M. Grygiel (Ed.), *Contrast and Analogy in Language: Perspectives from Cognitive Linguistics*. John Benjamins.
- Höder, S. (2014). Constructing diasystems: Grammatical organisation in bilingual groups. In T. Åfarli & B. Mæhlum (Eds.), *The Sociolinguistics of Grammar* (pp. 137–152). John Benjamins. <https://doi.org/10.1075/slcs.154.07hod>
- Höder, S. (2018). Grammar is community-specific: Background and basic concepts of Diasystematic Construction Grammar. In H. Boas & S. Höder (Eds.), *Constructions in Contact Constructional Perspectives on Contact Phenomena in Germanic languages* (pp. 37–70). Benjamins. <https://doi.org/10.1075/cal.24.02hod>
- Jaeger, H., Steels, L., Baronchelli, A., Briscoe, T., Christiansen, M., Griffiths, T., Jäger, G., Kirby, S., Komarova, N., Peter, R., & Jochen, T. (2009). What can mathematical, computational, and robotic models tell us about the origins of syntax? In *Biological Foundations and Origin of Syntax*. The MIT Press. <https://doi.org/10.7551/mitpress/9780262013567.003.0018>
- Karjus, A., & Ehala, M. (2018). Testing an agent-based model of language choice on sociolinguistic survey data. *Language Dynamics and Change*, 8(2), 219–252. <https://doi.org/10.1163/22105832-00802004>
- Landsbergen, F. (2009). *Cultural Evolutionary Modeling of Patterns in Language Change: Exercises in Evolutionary Linguistics*. LOT.
- Landsbergen, F., Lachlan, R., ten Cate, C., & Verhagen, A. (2010). A cultural evolutionary model of patterns in semantic change. *Linguistics*, 48(2), 363. <https://doi.org/10.1515/ling.2010.012>
- Levshina, N. (2015). *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. John Benjamins. <https://doi.org/10.1075/z.195>
- Meyer, D., Zeileis, A., & Hornik, K. (2020). *vcd: Visualizing Categorical Data* (Version 1.4-6) [Computer software]. <https://cran.r-project.org/web/packages/vcd/index.html>
- Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & Baayen, H. (2002). Experiences from the Spoken Dutch corpus project. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, 340–347.
- Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013a). *SoNaR User Documentation (version 1.0.4)*. [https://ticlops.uvt.nl/SoNaR\\_end-user\\_documentation\\_v.1.0.4.pdf](https://ticlops.uvt.nl/SoNaR_end-user_documentation_v.1.0.4.pdf)
- Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013b). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odiijk (Eds.), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing* (pp. 219–247). Springer. [https://doi.org/10.1007/978-3-642-30910-6\\_13](https://doi.org/10.1007/978-3-642-30910-6_13)


- Perek, F., & Goldberg, A. E. (2015). Generalizing beyond the input: The functions of the constructions matter. *Journal of Memory and Language*, 84, 108–127. <https://doi.org/10.1016/j.jml.2015.04.006>
- Pérez-Martín, A. M. (2007). Pluralización de *había* en el habla de El Hierro: Datos cuantitativos [Pluralization of *había* in the speech of El Hierro: Quantitative data]. *Revista de Filología de La Universidad de La Laguna*, 25, 505–513.
- Phan, D., & Varenne, F. (2010). Agent-based models and simulations in economics and social sciences. *Journal of Artificial Societies and Social Simulation*, 13(4), 1532. <https://doi.org/10.18564/jasss.1532>
- Pijpops, D. (2019). *How, Why and Where Does Argument Structure Vary? A Usage-Based Investigation into the Dutch Transitive-Prepositional Alternation* [Doctoral dissertation, University of Leuven]. LIRIAS @ KU Leuven. <https://lirias.kuleuven.be/2815151>
- Pijpops, D., & Beuls, K. (2015). Agent-gebaseerde modellering in de historische taalkunde. Een model van regularisatiedruk op de Nederlandse werkwoorden [Agent-based modelling in historical linguistics: A model of the regularization pressure on Dutch verbs]. *Handelingen Der Koninklijke Zuid-Nederlandse Maatschappij Voor Taal- En Letterkunde En Geschiedenis*, 69, 5–23.
- Pijpops, D., Beuls, K., & Van de Velde, F. (2015). The rise of the verbal weak inflection in Germanic: An agent-based model. *Computational Linguistics in the Netherlands Journal*, 5, 81–102.
- Pijpops, D., De Smet, I., & Van de Velde, F. (2018). Constructional contamination in morphology and syntax: Four case studies. *Constructions and Frames*, 10(2), 269–305. <https://doi.org/10.1075/cf.00021.pij>
- Pijpops, D., & Van de Velde, F. (2015). Ethnolect speakers and Dutch partitive adjectival inflection: A corpus analysis. *Taal En Tongval*, 67(2), 343–371. <https://doi.org/10.5117/TET2015.2.PIJP>
- Pijpops, D., & Van de Velde, F. (2016). Constructional contamination: How does it work and how do we measure it? *Folia Linguistica*, 50(2), 543–581. <https://doi.org/10.1515/flin-2016-0020>
- Pijpops, D., & Van de Velde, F. (2018). A multivariate analysis of the partitive genitive in Dutch: Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory*, 14(1), 99–131. <https://doi.org/10.1515/cllt-2013-0027>
- Plevoets, K. (2008). *Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands* [Between language for speaking and standard language. A corpus-based study to the situational, regional and social diffusion of a number of morphosyntactic features of spoken Belgian Dutch] [Doctoral dissertation, University of Leuven]. LIRIAS @ KU Leuven. <https://lirias.kuleuven.be/1821028>
- Ruette, T. (2012). *Aggregating Lexical Variation: Towards Large-scale Lexical Lectometry* [Doctoral dissertation, University of Leuven]. LIRIAS @ KU Leuven. <https://lirias.kuleuven.be/1821265>
- Ruette, T., Ehret, K., & Szmrecsanyi, B. (2016). A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics*, 21(1), 48–79. <https://doi.org/10.1075/ijcl.21.1.03rue>
- Sonderegger, M., Wagner, M., & Torreira, F. (2018). *Quantitative Methods for Linguistic Data*. <http://people.linguistics.mcgill.ca/~morgan/book/>

- Speelman, Dirk. (2014). Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In D. Glynn & J.A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (pp. 487–533). John Benjamins. <https://doi.org/10.1075/hcp.43.18spe>
- Speelman, D., Heylen, K., & Geeraerts, D. (2018). *Mixed-Effects Regression Models in Linguistics*. Springer. <https://doi.org/10.1007/978-3-319-69830-4>
- Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4), 339–356. <https://doi.org/10.1016/j.plrev.2011.10.014>
- Steels, L. (2000). Language as a complex adaptive system. In M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J.J. Merelo, & H.-P. Schwefel (Eds.), *Proceedings of PPSN VI: Lecture Notes in Computer Science* (pp. 17–26). Springer. [https://doi.org/10.1007/3-540-45356-3\\_2](https://doi.org/10.1007/3-540-45356-3_2)
- Tremblay, A., & Baayen, R.H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 151–173). Continuum.
- Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61(2), 569–613. <https://doi.org/10.1111/j.1467-9922.2010.00622.x>
- van Agtmaal-Wobma, E., Harmsen, C., Dal, L., & Poulain, M. (2007). *Belgen in Nederland en Nederlanders in België* [Belgians in the Netherlands and Dutchmen in Belgium]. Centraal Bureau voor Statistiek (CBS). <https://www.cbs.nl/-/media/imported/documents/2008/02/2007-k4-b15-p47-art.pdf>
- van den Toorn, M.C. (1977). *Nederlandse Grammatica* [Dutch Grammar] (5th ed.). Wolters-Noordhoff.
- van der Horst, J. (2008). *Geschiedenis van de Nederlandse syntaxis* [History of Dutch syntax]. Universitaire Pers Leuven.
- van Eerten, L. (2007). Over het Corpus Gesproken Nederlands [About the Corpus of Spoken Dutch]. *Nederlandse Taalkunde*, 12(3), 194–215.
- Van Rossum, G., & Drake, F.L. (2009). *Python 3 Reference Manual*. CreateSpace.
- van Trijp, R., & Steels, L. (2012). Multilevel alignment maintains language systematicity. *Advances in Complex Systems*, 15(3–4). <https://doi.org/10.1142/S0219525912500397>
- Wellens, P. (2012). *Adaptive Strategies in the Emergence of Lexical Systems*. Dissertation Vrije Universiteit Brussel.
- Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics*, 1, 243–264. <https://doi.org/10.1146/annurev-linguist-030514-124930>

## Appendices

The appendices are available from <https://doi.org/10.1075/ijcl.20040.pij.additional>

## Address for correspondence

Dirk Pijpops  
Department of Modern Languages  
University of Liège  
Place Cockerill 3-5  
Liège, 4000  
Belgium  
dirk.pijpops@uliege.be  
 <https://orcid.org/0000-0002-3820-8099>

## Publication history

Date received: 31 March 2020  
Date accepted: 12 April 2022  
Published online: 13 June 2022