# Multiple Correspondence Analysis as Heuristic Tool to Unveil Confounding Variables in Corpus Linguistics

Jose Tummers[1,2], Dirk Speelman[2], Dirk Geeraerts[2]

[1] KHLeuven – B-3001 Leuven – Belgium
[2] KULeuven, RU Quantitative Lexicology and Variational Linguistics – B-3000 Leuven – Belgium

**Abstract**

Corpus linguistic research relies on corpora which generally display an unbalanced structure. We will discuss a potential corollary of this biased structure which is rarely accounted for in (corpus) linguistics, namely confounding variables. These are variables increasing, diminishing or reversing an explanatory variable's marginal effect compared to its conditional effect. Analyzing four instances of confounding in a variational case study governed by a series of categorical explanatory variables, we will argue that these latent confounders can be unveiled modeling the co-occurrence patterns of the explanatory variables by means of a multiple correspondence analysis.

**Keywords:** corpus linguistics, confounding variables, multiple correspondence analysis, Simpson's paradox

## 1. Introduction

In observational studies, which prevail in corpus linguistics, variables are not randomly assigned to treatments as in experimental studies. This generally results in uncontrolled data sets with intertwined explanatory variables potentially confounding each other's effect on the response variable. Based on a variational study where a linguistic alternation is governed by a series of categorical explanatory variables, we will discuss four instances where confounders alter a variable's effect resulting in contradictory results between the bivariate and multivariate analyses. Given the complexity of the linguistic alternation under investigation, we will argue that an exploratory description of the mutual associations between the explanatory variables by means of a multiple correspondence analysis allows us to unveil the confounders causing the spurious effects in the bivariate analyses.

The remainder of this paper is organized as follows. First, we will outline the phenomenon of confounding variables, which is hardly (explicitly) addressed in (corpus) linguistics (section 2). Next, we will briefly sketch the alternation between two inflectional variants of the attributive adjective in Dutch, the case study used to illustrate the incidence of confounding in corpus linguistics (section 3). In section 4, we will argue for the use of multiple correspondence analysis to identify potential confounding variables causing spurious effects in the bivariate analyses. This paper will be closed by formulating three methodological conclusions.

## 2. Confounding variables

Labov (1972) distinguishes four data gathering techniques for linguistic research: introspection, inquiries, experimentation and observation. In observational studies, the product of spontaneous language use is analyzed (Tummers *et al.*, 2005). This is the typical research domain of corpus linguistics, where language phenomena are studied in their natural environment to identify the variables determining their use.

Language corpora are designed to reflect the language use of a linguistic community in (a) given situation(s). The composition of corpora entails that the multitude of potential explanatory variables can hardly be controlled for, especially when it concerns constructions and alternations governed by a complex network of explanatory variables.[1] The resulting biased structure of most language corpora is an important difference with experimental research where (i) respondents are randomly assigned to the experimental or control groups, and (ii) potentially confounding variables are controlled for by the balanced structure of the experiment. In practice, corpus studies generally precede experimental research to model the effect of and interactions between explanatory variables (Grondelaers & Speelman, 2007). Next, this observational model is used for the experimental design to control the impact of potentially confounding variables on the effect of the explanatory variable at stake (Grondelaers *et al.*, 2009), since not only subject-related variables but also linguistic variables are controlled for in (psycho)linguistic experiments.[2]

In the present paper, we will address a problem that (can) occur(s) in corpus-based studies as corollary of the uncontrolled and biased nature of corpus data, namely confounding variables affecting the effect of an explanatory variable on the response variable. This phenomenon is often denoted as Simpson's paradox. It is defined as the reversal of an explanatory variable's effect on the response variable by a confounding variable (Agresti, 2007; Pearl, 2000: chap. 6; Schield, 1999). As a result, the marginal effect observed in the bivariate analysis and the conditional effect observed in the multivariate analysis are opposed. Tu *et al.* (2008) extend this narrow definition of Simpson's paradox to cases where the marginal effect is diminished or enhanced without being reversed. We will use the term confounding (variables) to denote instances of confounding according to the broad definition of Tu *et al.* (2008).

Although the phenomenon of confounding variables is well-known in epidemiological studies (Reintjes *et al.*, 2000; Tu *et al.*, 2008), economy/econometrics (Lipovetsky & Conklin, 2006), management studies (Curley & Brown, 2000), social sciences (Nurmi, 1997), and judicial studies (Doob, 2007), it is hardly ever (explicitly) raised in (corpus) linguistics. This is surprising since "issues of confounding will (or should) invariably arise when using non-experimental data" (Greenland *et al.*, 1999: 29) and yield "controversial and contradictory results" (Tu *et al.*, 2008: 8). Moreover, when observational data are used to infer causal relations, spurious effects caused by confounding have to be precluded from the final model to avoid that Simpson's Paradox is "used to argue that induction is impossible in observational studies" (Schield, 1999: 106).[3]

1 For an account of the complexity of corpus data and its implications, we refer to Heylen *et al.* (2008).

2 In the statistical analysis, both sources of variation are known as F1 and F2 respectively (Rietveld & van Hout, 2005).

3 For a thorough discussion of the relation between causation and confounding, we refer to Greenland *et al.* (1999) and Pearl (2000: chap. 6), where this subject is addressed for observational as well as experimental studies.

In the remainder of this paper, we will outline a corpus-based account of an inflectional alternation in Dutch where spurious effects occur in the bivariate analyses as a result of confounding variables.

## 3. Dutch adjectival inflection: a complex alternation with confounding variables

### *3.1. Dutch adjectival inflection*

In Dutch, attributive adjectives in definite NPs with a singular neuter head noun exhibit a complex morphological alternation between the unmarked inflected adjective and the marked uninflected adjective, as illustrated by the following examples. The inflected adjective is composed by suffixation of the morpheme *-e* ([ə]).

1. het vriendelijk-e kind
   the friendly-INFL child
2. (2)het vriendelijk-ø kind
   the friendly-ZERO child

The use of the marked uninflected adjective is governed by an intricate network of variables (Haeseryn *et al.*, 1997; Rooij, 1980a, 1980b; Lebrun & Schurmans-Swillen, 1966; Tummers, 2005). These explanatory variables can be structural, discourse-related and lectal in nature. Table 1 schematizes the effect of the explanatory variables as discussed in literature.

| Type | Variable value inflected adjective | Variable | Variable value uninflected adjective |
|---|---|---|---|
| structural | definite article, demonstrative determiner, genitive | POS determiner | possessive determiner |
| | positive | gradation A | comparative |
| | no | diminutive N | yes |
| | no | lexical unity AN | yes |
| | qualifying A | semantic category A | relational A |
| discourse-related | consonant | onset N | vowel |
| | monosyllabic, bisyllabic | length A | plurisyllabic |
| | stressed | stress final syllable A | unstressed |
| | stressed | stress first syllable N | unstressed |
| lectal | Netherlandic Dutch | national variety | Belgian Dutch |
| | formal, unmarked | register belgian.dutch | informal |
| | unmarked | register netherlandic.dutch | (very) formal |

*Table 1: Determinants of both inflectional alternatives of the attributive adjective in definite NPs with singular neuter head noun*

For a detailed account of the notion of causation in observational and experimental studies in social sciences, we refer to Goldthorpe (2001).

We have studied this inflectional alternation in spoken Dutch using the Corpus of Spoken Dutch (*Corpus Gesproken Nederlands* (CGN); Oostdijk, 2000). This is a 4M word corpus of spoken Dutch, containing data from Belgian and Netherlandic Dutch, both national varieties of Dutch, and various registers ranging from highly informal (viz. colloquial speech) to highly formal (viz. prepared speeches in parliament). The distribution of both inflectional variants in the corpus is summarized in table 2. The systemically marked status of the uninflected adjective is clearly confirmed by the frequency data.

| | **n** | **%** |
|---|---|---|
| **inflected** | 3,810 | 76.75 |
| **uninflected** | 1,154 | 23.25 |
| Σ | **4,964** | **100.00** |

*Table 2: Distribution in CGN of both inflectional alternatives of the attributive adjective in definite NPs with singular neuter head noun*

### *3.2. Confounding and spurious effects in adjectival inflection*

The analysis of the inflectional alternation consists of two stages. In the first stage, we have performed bivariate analyses. Focusing on the operationalization of the explanatory variables, the marginal effect on the response variable has been computed according to various operationalizations in order to optimize the impact of the explanatory variable and the linguistic interpretation. In the second stage, a multivariate model has been constructed integrating all potential explanatory variables. Performing a binary logistic regression analysis, the impact of each explanatory variable as well as its significance is modeled controlling for all other explanatory variables in the model. This model represents the conditional effect of each explanatory variable. The comparison of the marginal and the conditional effect of the potential explanatory variables revealed two strict instances of confounding, viz. a reversal of the effect, and two weak instances, viz. a reversal of the significance without alteration of the direction of the effect.

The explanatory variables' effects in the bivariate and the multivariate models will be compared by means of the odds ratio (OR) and its 95% CI (Agresti, 2007: 28-33). The OR compares the odds $^{\text{uninflected}}/_{\text{inflected}}$, viz. $^{\text{success}}/_{\text{fail}}$ in the present analysis. The explanatory variables are recoded in dummy variables by means of reference coding (Davis, 2010).[4] The significance of the explanatory variable's effect is expressed by means of the 95% CI, which belongs to the interval between 0 and +Inf with 1 as pivot:

- A score of 1 identifies the absence of any relation between the explanatory variable and the response variable.
- A 95% CI with an upper boundary smaller than 1 identifies a significant negative relation: the odds uninflected/inflected significantly diminish for the marked value of the explanatory variable compared to the reference value.

4 The systemically unmarked value of the explanatory variable generally functions as reference value.

- A 95% CI with a lower boundary greater than 1 identifies a significant positive relation: the odds uninflected/inflected significantly increase for the marked value of the explanatory variable compared to the reference value.

Table 3 summarizes the coding of the explanatory variables at the outcome of the bivariate analyses. This table contains two complex variables: 'a.len.pros' and 'reg.inf.index'. These variables merge two variables, the former in order to avoid multicollinearity, the latter to deal with an interaction. The variable 'a.len.pros' combines the length of the adjective (values: 'one' vs. 'two' vs. 'more' syllables) and the prosodic pattern of the rightmost metric unit of the adjective (values: 's', 'sw', 'sww')[5], which are strongly correlated. The other complex variable combines the interacting lectal variables, viz. the regional variety of Dutch (values: 'nl' and 'bel' for Netherlandic and Belgian Dutch respectively) and the informality index of the register[6] (values: '0' vs. '1' vs. '2' vs. '3', ranging from very formal to very informal). This complex variable has been created in order to compare the bivariate and multivariate analyses in table 4. The lexical collocability between the A and the N is measured by computing the log likelihood ratio of the AN pair (Dunning, 1993). AN pairs without significant collocability are the reference value; AN pairs displaying a significant degree of collocability are grouped in 4 quartiles.

| **Variable** | **Reference value** | **Values** |
|---|---|---|
| det.cat (POS determiner) | def.article | possessive, demonstrative, genitive |
| a.comp (comparative A) | no | yes |
| n.inf (nominalized infinitive as N) | no | yes |
| n.dim (use diminutive N) | no | yes |
| n.gender (gender N) | neuter | bigeneric |
| a.sonor (sonority final vowel A) | no | yes |
| n.onset (onset N) | consonant | vowel |
| a.len.pros (length & prosodic pattern rightmost metric unit A) | one.s | two.s, two.sw, more.s, more.sw, more.sww |
| n.pros (prosodic pattern leftmost syllable N) | s | w |
| reg.inf.index (region & degree of informality) | nl.0 | nl.1, nl.2, nl.3, bel.0, bel.1, bel.2, bel.3 |
| lex.col (lexical collocability AN) | no | q.1, q.2, q.3, q.4 |
| a.rel (relational A) | no | yes |
| a.deriv (derivationally complex A) | no | yes |

*Table 3: Operationalization explanatory variables*

5 In prosodic patterns, 's' stand for a stressed syllable and 'w' for an unstressed or weak syllable. For instance, the code 'two.sw' stands for a bisyllabic adjective with 'sw' as prosodic pattern.

6 The informality index of the register is computed by combining the values of the three stylistic dimensions in the CGN corpus, namely interaction (opposing dialogue/multilogue to monologue), speaker preparation (opposing spontaneous to prepared speech), and audience (opposing private to public). Combining these dimensions, the lowest (i.e. most formal) informality degree of '0' refers to prepared, public monologues, whereas the highest informality degree of '3' identifies spontaneous, private dialogues/multilogues.

Table 4 presents the ORs for the bivariate and the multivariate analyses. The instances of confounding are marked in bold case. The code '/' for the OR and its 95% CI in the multivariate analysis indicates that the variable has not been included in the stepwise regression model for not being significant.

In order to assess the regression model[7], we will briefly discuss the model statistics. The model significantly reduces the total variance (measured by using Akaike Information Criterion): $\text{variance}_{\text{EXPLAINED}} = \text{variance}_{\text{TOTAL}} - \text{variance}_{\text{RESIDUAL}} = 5{,}383.4 - 3{,}779.7 = 1{,}603.7$, df = 25, $p < 0.00001$.[8] This result is corroborated by the model's predictive power: $c = 0.8555$.[9]

| **Explanatory variable** | | **Bivariate analysis** | | **Multivariate analysis** | |
|---|---|---|---|---|---|
| **Reference** | **Value** | **OR** | **95% CI** | **OR** | **95% CI** |
| det.cat:def.article | det. cat:demonstrative | 0.6854 | [0.5458,0.8607] | 0.8652 | [0.6547,1.1435] |
| | **det.cat:possessive** | **0.8624** | **[0.6960,1.0687]** | **2.0266** | **[1.5512,2.6478]** |
| | det.cat:genitive | 0.8325 | [0.2756,2.5145] | 1.5681 | [0.4157,5.9145] |
| a.comp:no | a.comp:yes | 0.9707 | [0.5437,1.7328] | / | / |
| n.dim:no | **n.dim:yes** | **0.4133** | **[0.2928,0.5834]** | / | / |
| a.sonor:no | **a.sonor:yes** | **0.5009** | **[0.4384,0.5724]** | / | / |
| n.inf:no | n.inf:yes | 2.595 | [1.7091,3.9399] | 4.3745 | [2.6062,7.3425] |
| n.gender:neuter | n.gender:bigen | 0.4273 | [0.2823,0.6469] | 0.6085 | [0.3807,0.9725] |
| n.onset:consonant | n.onset:vowel | 1.9193 | [1.6019,2.2996] | 1.7113 | [1.3516,2.1668] |
| a.len.pros:one.s | a.len.pros:two.s | 1.3448 | [0.9988,1.8108] | 1.3675 | [0.9337,2.0027] |
| | a.len.pros:two.sw | 3.5117 | [2.7889,4.4217] | 3.3048 | [2.3829,4.5835] |
| | a.len.pros:more.s | 5.4554 | [4.4165,6.7387] | 2.8545 | [2.0852,3.9075] |
| | a.len.pros:more.sw | 6.3706 | [5.0340,8.0620] | 6.7429 | [4.7494,9.5732] |
| | a.len.pros:meer. sww | 18.6793 | [14.1886,24.5913] | 18.406 | [12.645,26.792] |
| n.pros:s | n.pros:w | 1.4269 | [1.2412,1.6403] | 1.5099 | [1.2691,1.7965] |
| reg.inf.index:nl.0 | reg.inf.index:nl.1 | 2.3672 | [0.7443,7.5284] | 3.2352 | [0.9678,10.814] |
| | reg.inf.index:nl.2 | 3.9862 | [3.0394,5.2278] | 2.882 | [2.068,4.0164] |
| | reg.inf.index:nl.3 | 1.2977 | [1.0016,1.6812] | 2.3731 | [1.7366,3.2427] |
| | reg.inf.index:bel.0 | 1.4054 | [1.1350,1.7401] | 1.3033 | [1.014,1.6751] |
| | reg.inf.index:bel.1 | 3.4561 | [2.6564,4.4965] | 2.6767 | [1.9352,3.7023] |
| | reg.inf.index:bel.2 | 4.7467 | [3.6647,6.1481] | 5.1851 | [3.7902,7.0934] |

7 The regression statistics have been computed by means of the `Design` library in R (Harrel, 2001).

8 When testing for multicollinearity, only two variables slightly exceed the lower variance inflation (VIF) threshold of 2.5, namely 'a.len.pros:more.s' (VIF = 2.8780) and 'a.len.pros:more.sw' (VIF = 2.7575). We do not consider these weak crossings of the threshold a treat for the stability of the regression model.

9 The *c* statistic computes the area under the *Receiver Operating Characteristic* curve. This curve plots the sensitivity (correctly predicted successes) of a model against 1 - its specificity (correctly predicted fails). When interpreting the *c* statistic ($c \in [0.5,1.0]$), $0.8 \leq c < 0.9$ is considered to be indicative of a very good model.

| | | | | | |
|---|---|---|---|---|---|
| | reg.inf.index:bel.3 | 5.6391 | [4.2373,7.5046] | 9.3166 | [6.5159,13.321] |
| lex.col:no | **lex.col:q.1** | **0.6152** | **[0.4494,0.8420]** | **1.2266** | **[0.8514,1.7671]** |
| | lex.col:q.2 | 1.5807 | [1.2403,2.0146] | 1.8624 | [1.4048,2.4692] |
| | lex.col:q.3 | 2.8088 | [2.2571,3.4954] | 3.7210 | [2.8252,4.9008] |
| | lex.col:q.4 | 8.1688 | [6.5724,10.1528] | 9.0087 | [6.8439,11.858] |
| a.rel:no | a.rel:yes | 4.1055 | [3.5754,4.7143] | 2.4704 | [2.0384,2.994] |
| a.deriv:no | a.deriv:yes | 5.9274 | [4.9269,7.1311] | 1.416 | [1.0334,1.9404] |

*Table 4: Comparison of effect values explanatory variables in bivariate and multivariate analyses*

The following conclusions can be drawn from the comparison between the bivariate and the multivariate analyses. Firstly, although most effects as well as their (non-)significance are confirmed, the conditional effect often adjusts the size of the marginal effect. Secondly, in two cases, we observe a reversal of the effect. (i) While the OR of the bivariate analysis suggests a non-significant decreasing effect of the possessive determiner on the odds $^{\text{uninflected}}/_{\text{inflected}}$, the OR of the multivariate analysis indicates a significant increase of these odds, as suggested in the literature overview in table 1. (ii) For the lowest degree of lexical collocability ('lex.col:q.1'), the significant decrease of the odds $^{\text{uninflected}}/_{\text{inflected}}$ in the bivariate analysis is turned into a non-significant increase in the multivariate analysis. Thirdly, two parameters exhibit a significant effect in the bivariate analysis which disappears in the multidimensional model, namely the sonority of the nominal onset ('n.sonor') and the use of a diminutive noun ('n.dim'). The following step is to explain these discrepancies between the marginal effects in the bivariate analysis and the conditional effects in the multivariate analysis.

## 4. Multiple correspondence analysis as heuristic tool to identify confounding variables

Most examples of confounding in literature mention an explanatory variable whose effect is altered by one confounder. In these cases, a stratum and/or multidimensional analysis exposes the effect of the confounding variable due to a distributional bias. Schield (1999) proposes a formal criterion to identify one-by-one potentially confounding variables stating the minimal effect size of a potential confounder.[10]

The case study of interest and, by extension, most corpus linguistic studies deal with a multitude of potential explanatory variables. This implies that the potential confounding variables have not to be identified in isolation according to the above mentioned criterion but in interaction with other potential confounding variables, i.e. all potential explanatory variables. As a consequence, the uncontrolled nature of corpus data requires next to multivariate modeling a description of the (complex relations between the explanatory variables in the) data matrix. Since all explanatory variables in the case study are categorical in nature, a multiple correspondence analysis (MCA) will be performed to model the relations between the potential explanatory variables.

10 Schield (1999: 109) presents the following equation to be checked for every potential confounder: $[P(D|E) - P(D|\sim E)] \geq [P(E|A) - P(E|\sim A)]$ (where D stands for the response variable, E for the explanatory variable, A for the potential confounder, and ~X for the complement of X).

MCA (Greenacre, 1984, 2006, 2007) is an exploratory technique to identify and visualize the relation(s) between variable values. The information provided by a MCA is twofold. First, the original *n* dimensions are reduced to latent dimensions. These dimensions are computed according to their contribution to the global $\chi^2$-statistic for the complete data matrix. Only a limited number of latent dimensions are used for interpretation based on the inertia they explain. Next, the variable values are assigned a position with respect to these latent dimensions which are the axes of two-dimensional (or three-dimensional) plots. In those plots, variable values assigned to the same quadrant are associated. Taking into account our aim, viz. explaining the origin of four instances of confounding, the relative positions of the variables values rather than the interpretation of the latent dimensions are our main concern. The MCA has been performed by means of the `ca` library in R (Nenadic & Greenacre, 2007).

The input for the MCA is the data matrix as used for the logistic regression in table 4, with omission of the response variable. The analysis is based on the Burt matrix with an adjustment of the inertias, the so-called "adjusted analysis" (Greenacre, 2006, 2007). Table 5 summarizes the principal inertias (eigenvalues), the percentages and cumulative percentages for all dimensions of the data matrix, and presents a scree plot of the principal inertias. These figures motivate an analysis confined to the first two dimensions: the so-called elbow in the scree plot occurs after dimension 2 and the first two dimensions are the only dimensions explaining more than 5 percent of the inertia.

| Dim | Value | % | Cum % | Scree plot |
|---|---|---|---|---|
| 1 | 0.018085 | 61.7 | 61.7 | ********** |
| 2 | 0.001562 | 5.3 | 67.1 | ** |
| 3 | 0.000824 | 2.8 | 69.9 | * |
| 4 | 0.000441 | 1.5 | 71.4 | * |
| 5 | 0.000270 | 0.9 | 72.3 | |
| 6 | 0.000176 | 0.6 | 72.9 | |
| 7 | 7.2e-050 | 0.2 | 73.1 | |
| 8 | 5e-05000 | 0.2 | 73.3 | |
| 9 | 3.2e-050 | 0.1 | 73.4 | |
| 10 | 9e-06000 | 0.0 | 73.5 | |
| 11 | 2e-06000 | 0.0 | 73.5 | |
| 12 | 00000000 | 0.0 | 73.5 | |
| **Total** | **0.029300** | | | |

*Table 5: Summary of adjusted MCA of data matrix after omission of response variable*

The results of the MCA are visualized in figure 1 (symmetric map). Variable values in the same quadrant are closely associated in the data matrix, viz. they co-occur more often than expected by mere chance. In other words, the MCA plot can be considered a visualization of co-occurrence patterns between categorical variable values. It has to be noticed that in the remainder of this contribution we use the notions of co-occurrence and association patterns to

refer to patterns of variable values in the data matrix and not to lexical patterns in language use, as it is generally the case in corpus linguistics.

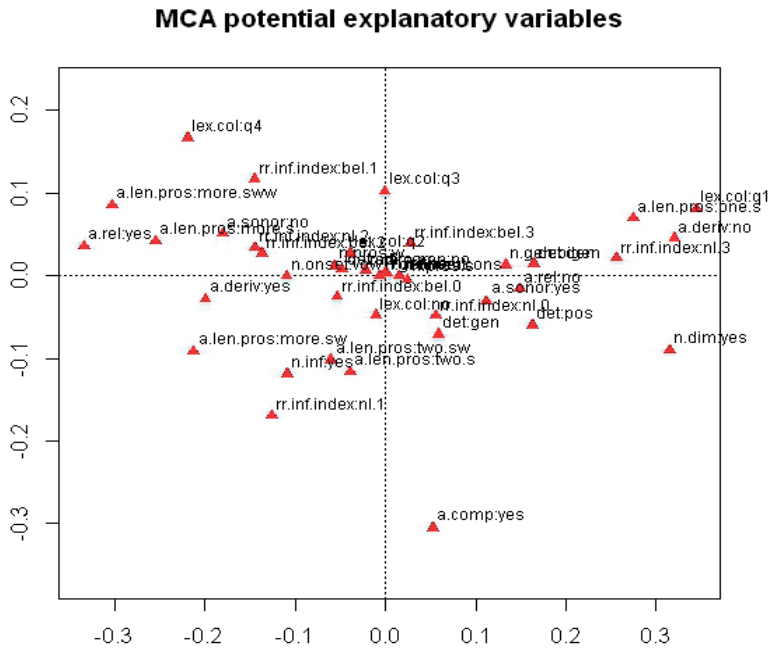


*Figure 1: Plot of first two dimensions adjusted MCA of potential explanatory variables*

In the remainder of this section, we will focus on the quadrants in figure 1 where the variable values involved in the four cases of confounding are situated. For every case of confounding, we will compare two plots, the left-hand plot representing the quadrant where the reference value of the variable under interest (denominator of the OR in table 4) is situated, the right-hand plot reproducing the quadrant where the variable value with the spurious effect in the bivariate analysis (nominator of the OR in table 4) is situated. The variable values are identified by means of 4 symbols: the variable values at stake are represented by an asterisk ('*'), the (red) triangles point down identify the variable values significantly increasing the odds $^{\text{uninflected}}/_{\text{inflected}}$ in the multivariate model, the (green) triangles point up identify the variable values significantly decreasing the odds $^{\text{uninflected}}/_{\text{inflected}}$ in the multivariate model, and the (black) dots identify the variable values with no significant impact on the odds $^{\text{uninflected}}/_{\text{inflected}}$ in the multivariate model.

We will start by analyzing the co-occurrence patterns of the determiner values in the data matrix, to be more precise the co-occurrence patterns of the definite article (reference value to compute the OR) and the possessive determiner (variable value subject to reversal of effect). Figure 2 visualizes the co-occurrence patterns of both determiners.

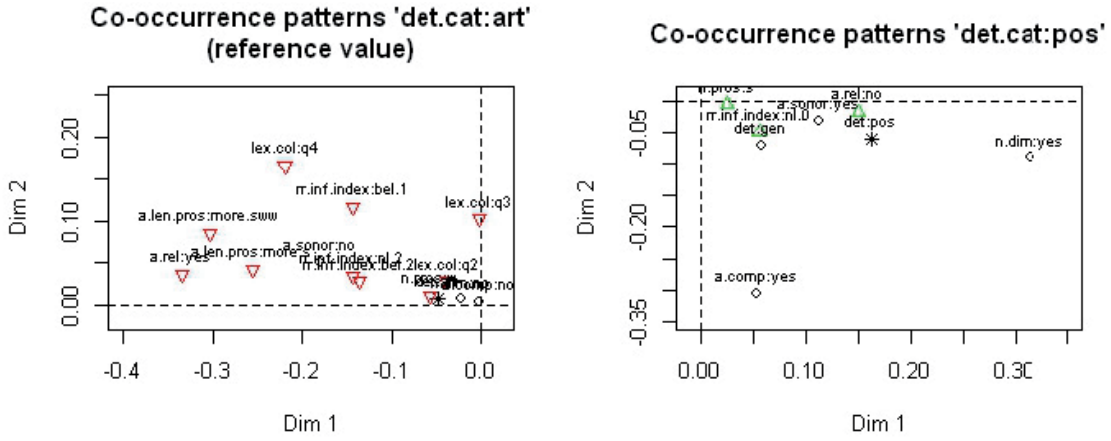


*Figure 2: Variable values co-occurring with the definite article and possessive determiner*

The data in the left-hand plot of figure 2 clearly show that the increasing effect of the definite article on the odds $^{\text{uninflected}}/_{\text{inflected}}$ in the bivariate analysis results from its association in the data matrix with mainly odds increasing variable values (red triangles point down). On the other hand, the possessive determiner's spurious decreasing effect on the odds $^{\text{uninflected}}/_{\text{inflected}}$ in the bivariate analysis is caused by its association with mainly odds decreasing variable values (green triangles top up in right-hand plot in figure 2). Since these multiple confounding variable values are controlled for in the multivariate analysis, the real effect of the possessive determiner on the adjectival inflection shows up.

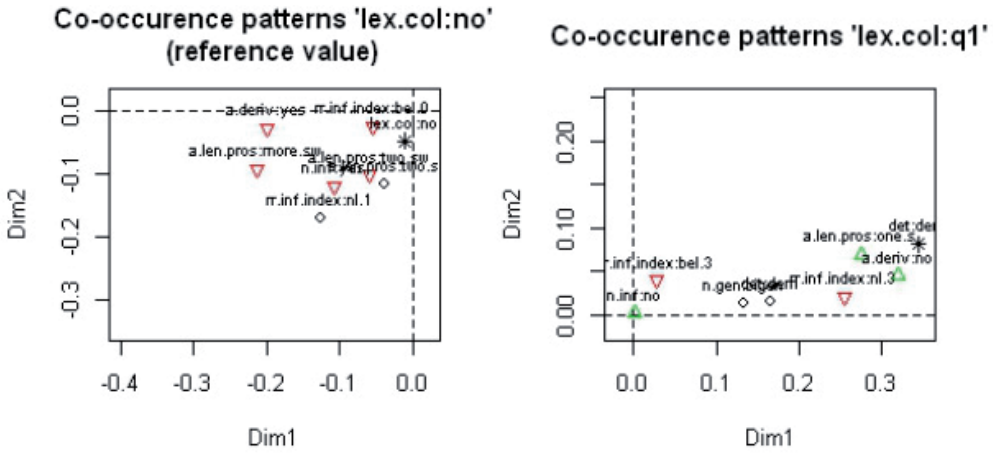


*Figure 3: Variable values co-occurring with 'lex.col:no' and 'lex.col:q.1'*

We will now consider the reversal of the effect of the lowest degree of lexical collocability ('lex.col:q.1'). The co-occurrence patterns in the data matrix of this value and its reference value ('lex.col:no') are displayed in figure 3. The increasing effect on the odds $^{\text{uninflected}}/_{\text{inflected}}$ of the lowest degree of collocability in the bivariate analysis is reversed in the multivariate analysis. The reference value almost exclusively co-occurs with variable values increasing the odds, as shown by the overwhelming presence of red point down triangles in the left-hand plot in figure 3. Although the situation for the lowest degree of lexical collocability is less clear-cut,

this variable value exhibits co-occurrence patterns with mainly variable values decreasing the odds $^{\text{uninflected}}/_{\text{inflected}}$.

The two remaining variables, 'a.sonor' and 'n.dim', will be treated together since they exhibit the same difference between the bivariate and the multivariate analyses, and their respective values are situated in the same quadrants of figure 1. Both variables display a significant odds decreasing effect in the bivariate analyses, whereas they have not been included in the stepwise regression model for not being significant.

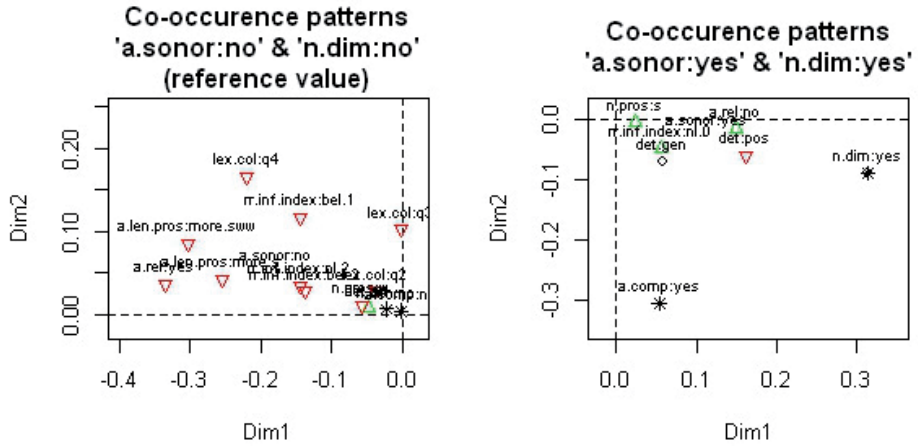


*Figure 4: Variable values co-occurring with values of 'a.sonor' and 'n.dim'*

For both variables, the reference value co-occurs with variable values increasing the odds $^{\text{uninflected}}/_{\text{inflected}}$, as indicated by the red top down arrows in the left-hand plot of figure 4. The other values are associated to variable values decreasing the odds $^{\text{uninflected}}/_{\text{inflected}}$, as identified by the majority of green top up triangles in the right-hand plot. As a result, the bivariate analysis yields a spurious significant effect which disappears in the multivariate analysis when confounders are controlled for.

In the four examples we have discussed in this section, the same pattern shows up. The real effect of the explanatory variable value cannot be deduced from the bivariate analysis as a result of the variable's association to explanatory variables with an opposite effect on the response variable. These association patterns result from the uncontrolled and biased structure of a corpus, where a balanced organization of explanatory parameters is quasi impossible to achieve.

## 5. Discussion and conclusion

We have studied four instances of confounding variables affecting the effect of explanatory variables in a corpus-based analysis of Dutch inflectional variation. We did not only observe a reversal of the effect of two explanatory variables, but also two variables with a significant effect in the bivariate analysis being no longer significant in the multivariate analysis. In all four cases, the real effect of the explanatory variable was altered in the bivariate analysis by its association patterns in the data matrix. Due to their association with variable values with an effect opposite to theirs, the impact of these variable values on the response variable was reversed or at least altered when the results of the bivariate and multivariate analysis were compared.

The analysis of this case study supports the following three methodological claims for corpus linguistic research in general and variationist research in particular. Firstly, due to the very nature of the data analyzed and the methodology applied, corpus linguistic research requires multivariate analyses to model the actual effect of explanatory variables on the response variable. This does not mean that we question the importance of bivariate analyses to operationalize the effect of potential explanatory variables, but that these bivariate analyses can only be the first research stage which has to be complemented by a multivariate analysis to avoid spurious effects of explanatory variables by controlling for potential confounding variables. In this respect, a clear methodological progress can be observed during the last decade (Gries, 2011), although some domains applying corpus linguistic methods still leave room for improvement (Römer & Wulf, 2010; Stefanowitsch, 2011).

Secondly, the use of more complex analytical techniques demands for a thorough exploration of the explanatory variables and their mutual associations in the data matrix. As shown in the present case study, hidden data matrix patterns can cause spurious effects in the bivariate analyses. These data matrix patterns are different from interaction effects and multicollinearity patterns, which are generally accounted for in multivariate studies. Different techniques, such as MCA for categorical explanatory variables, are needed to unveil potentially confounding associations in the data matrix.

Finally, corpus linguistics can methodologically still benefit from other disciplines where observational studies are used and which have a longer tradition of replication studies, such as epidemiology and econometrics. The present attempt to explore whether MCA can be used to unveil confounding variables of course needs to be falsified by other corpus linguistic and variationist case studies.

## References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Curley, S.P. & G.J. Browne (2000). "Normative and Descriptive Analyses of Simpson's Paradox in Decision Making". *Organizational Behavior and Human Decision Processes* 84(2): 308-333.

Davis, M. (2010). "Contrast coding in multiple regression analysis: Strengths, Weaknesses, and Utility of popular Coding Structures". *Journal of Data Science* 8: 61-73.

Doob, A. (2007). "The sentencing of aboriginal and non-aboriginal youth: Understanding local variation". *Canadian Journal Of Criminology And Criminal Justice* 49 (1): 109-123.

Dunning, T. (1993). "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics* 19(1): 61-74.

Goldthorpe, J. H. (2001). "Causation, Statistics, and Sociology". *European Sociological Review* 17(1): 1-20.

Greenacre, M. (1984). *Theory and Application of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (2006). "From Simple to Multiple Correspondence Analysis". In: M. Greenacre & J. Blasius (eds.), *Multiple Correspondence Analysis and Related Methods*, 41-77. London: Chapman.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. London: Chapman & Hall.

Greenland, S., J.M. Robins & J. Pearl. (1999). "Confounding and Collapsibility in Causal Inference". *Statistical Science* 14(1): 29-46.

Gries, S.Th. (2011). "Commentary". In: K. Allan & J. Robinson (eds.), *Current methods in historical semantics*, 184-195. Berlin & New York: Mouton de Gruyter.

Grondelaers, S. & D. Speelman (2007). "A variationist account of constituent ordering in presentative sentences in Belgian Dutch". *Corpus Linguistics and Linguistic Theory* 3: 161-193.

Grondelaers, S., D. Speelman, D. Drieghe, M. Brysbaert & D. Geeraerts (2009). "Introducing a new entity into discourse: Comprehension and production evidence for the status of Dutch *er* "there" as a higher-level expectancy monitor." *Acta Psychologica* 130: 153-160.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997). *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff Uitgevers – Deurne: Wolters Plantyn.

Harrell, F.E. (2001). *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*. New York: Springer.

Heylen, K., J. Tummers & D. Geeraerts (2008). "Methodological issues in corpus-based Cognitive Linguistics". In: G. Kristiansen & R. Dirven (eds.), *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems*, 91-128. Berlin: Mouton de Gruyter.

Labov, W. (1972). "Some principles of linguistic methodology". *Language in Society* 1: 97-120.

Lebrun, Y. & G. Schurmans-Swillen (1966). "Verbogen tegenover onverbogen adjectieven in de taal van de Zuidnederlandse dagbladpers". *Taal en Tongval* 18(1): 175-187.

Lipovetsky, S. & W.M. Conklin (2006). "Data aggregation and Simpson's paradox gauged by index numbers". *European Journal of Operational Research* 172: 334-351.

Nenadic, O. & M. Greenacre (2007). "Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The `ca` Package". *Journal of Statistical Software* 20(3). http://www.jstatsoft.org/.

Nurmi, H. (1997). "Voting paradoxes and referenda". *Social Choice and Welfare* 15(3): 333-350.

Oostdijk, N. (2000). "Het Corpus Gesproken Nederlands". *Nederlandse Taalkunde* 5(3): 280-284.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: CUP.

*Reintjes, R., A. de Boer A, W. van Pelt & J. Mintjes-de Groot (2000). "Simpson's paradox: an example from hospital epidemiology." Epidemiology* 11*:* 81-*83.*

Rietveld, T. & R. van Hout (2005). *Statistics in language research: Analysis of variance*. New York: Mouton de Gruyter.

Rooij, J. de (1980a). "Ons bruin(e) paard I". *Taal en Tongval* 32: 3-25.

Rooij, J. de (1980b). "Ons bruin(e) paard II". *Taal en Tongval* 32: 109-129.

Römer, U. & B. Wulf (2010). "Applying corpus methods to written academic texts: Explorations of MICUSP". *Journal of Writing Research* 2(2): 99-127.

Schield, M. (1999). "Simpson's paradox and Cornfield's conditions". *ASA-JSM, Proceedings of the Section of Statistical Education*: 106-111.

Stefanowitsch, A. (2011). "Cognitive linguistics meets the corpus". In: M. Brda, M. Fuchs & S. Gries (eds.), *Expanding cognitive linguistic horizons*, 257-288. Amsterdam: John Benjamins.

Tu, Y.-K., D. Gunnell & M. Gilthorpe (2008). "Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon - the reversal paradox". *Emerging Themes in Epidemiology* 5(2): 1-9.

Tummers, J. (2005) *Het naakte adjectief. Kwantitatief-empirisch onderzoek naar de adjectivische buigingsalternantie bij neutra*. PhD dissertation, KULeuven, Faculty of Arts.

Tummers, J., K. Heylen & D. Geeraerts (2005). "Usage-based approaches in Cognitive Linguistics: A technical state of the art". *Corpus Linguistics and Linguistic Theory* 1(2): 225-261.