Quantifying semantic effects. The impact of lexical collocations on the inflectional variation of Dutch attributive adjectives

Jose Tummers, Dirk Speelman, Dirk Geeraerts

KU Leuven – RU Quantitative Lexicology and Variational Linguistics
Blijde Inkomststraat 21 – 3000 Leuven – Belgium
{jose.tummers;dirk.speelman;dirk.geeraerts}@arts.kuleuven.ac.be

Abstract

This paper wants to verify to what extent semantic effects contribute to the alternation between inflected and uninflected adjectives in a definite NP with a singular neuter head noun as observed in the *Corpus Gesproken Nederlands*. Existing accounts invoke the non-compositional semantics of the adjective-noun combination to interpret this case of inflectional variation, generally using highly idiomatic examples. We will argue for a combinatorial operationalisation of the semantic effect, based on lexical collocations. The influence of lexical collocations on the inflectional alternation will be measured by means of the log-likelihood test. Finally, the actual impact of lexical collocations on the inflectional alternation will be quantified integrating the semantic parameter together with other potential explanatory factors in a logistic regression analysis.

Keywords: variational linguistics, adjectival inflection, Dutch, lexical collocation, log-likelihood ratio, logistic regression.

1. Introduction

Dutch attributive adjectives are always inflected, taking an *-e* suffix (*het mooi-e paard* 'the beautiful horse'), except in indefinite NPs with a singular neuter head noun, where the adjective remains uninflected (*een mooi-∅*[1] *paard* 'a beautiful horse').

A special case within this inflectional system are definite NPs with a singular neuter head noun, which allow for an uninflected adjective next to the 'regular' inflected alternative, such as *het bijvoeglijk-∅ naamwoord* (lit. 'the attributive noun', hence 'the adjective') and *het koninklijk-∅ besluit* ('the royal decree'). Although a number of parameters influence this alternation, the most quoted explanation regards the semantics of the adjective-noun combination: whenever the adjective-noun combination has a non-compositional semantics, the adjective is said to remain uninflected. The present contribution aims at implementing this semantic parameter in order to test its empirical validity.

This paper is further structured as follows. Section 2 discusses the variation observed in definite NPs with a singular neuter head, demonstrating the need for an empirical analysis. In section 3, we propose to make the semantic parameter operational using the log-likelihood test to measure the lexical attraction between the adjective and the noun. Next, the results of the empirical analysis will be presented and evaluated by integrating the semantic parameter

[1] We use '∅' to identify the uninflected adjective. We do not take any position with respect to the existence of a zero-morpheme. This discussion would amply exceed the scope of this paper.

in a multifactorial analysis. We will conclude by summarising the most important findings and raise some issues for further research.

2. Inflectional variation in definite NPs with a singular neuter head noun

Dutch adjectival inflection challenges the classical model of agreement, where features present in the agreement domain, i.e. the NP, spell out the inflectional suffix, viz. *-e* or *-∅*. Definite NPs with a singular neuter head noun allow the use of both suffixes, resulting in an intricate case of variation. Several studies show that this inflectional alternation is governed by a series of parameters pertaining to different linguistic domains such as phonology, morpho-syntax and lexical semantics as well as lectal factors such as register and the distinction between the two national varieties of Dutch (Lebrun and Schurmans-Swillen, 1966; de Rooij, 1980; Booij, 1992; ANS, 1997). We will briefly discuss the effects of the relevant parameters.

First, in NPs with a possessive pronoun as a determiner, such as *mijn sterk-∅ paard* ('my strong horse'), the uninflected adjective is more widespread than in sequences where the definite article *het* ('the') is used, such as *het sterk-∅ paard* ('the strong horse'). Second, rhythmic properties of the adjective may influence the inflection. The longer the adjective, the more natural the uninflected form seems to be, as in *het ongedifferentieerd-∅ corpus* ('the undifferentiated corpus'). This tendency is reinforced when the adjective ends in an unstressed or unvoiced syllable, such as *het ontzagwekkend-∅ paard* ('the awe-inspiring horse'). Third, semantically speaking, the use of the uninflected adjective can be observed when the adjective-noun sequence constitutes a semantic unity, such as *het Algemeen-∅ Nederlands* ('the standard variety of Dutch'). Fourth, from a lectal point of view, stylistic and geographical parameters modify the adjectival inflection. Stylistically, the uninflected alternative is considered to be more widespread in less formal contexts. Geographically, Belgian Dutch favours the uninflected adjective compared to Netherlandic Dutch, where the inflected counterpart is predominant.

In the rest of this paper, we will focus on the semantic parameter. Almost all scholars analysing the inflectional alternation cite this parameter and tend to consider it the most important explanatory factor to interpret the use of the uninflected adjective. As a matter of fact, several analyses even try to account for the use of the uninflected adjective invoking the sole semantic parameter (Honselaar, 1980; Odijk, 1992; Booij, 2002).

The semantic parameter points out that the use of the uninflected adjective identifies the NP as having non-fully compositional semantics. In such cases, the adjective-noun sequence constitutes a lexical unit that provides the name for a group of entities. The uninflected adjective *groot-∅* ('great') in *het groot-∅ seminarie* (lit. 'the great seminar') does not specify the referent of *seminarie* ('seminar') with respect to its size – as would the inflected adjective *groot-e* in *het groot-e seminarie* ('the great seminar, viz. the building'). Instead, the NP as a whole refers to an institution where boys are prepared for catholic priesthood. Booij (2002) uses the term "collocational idiom", emphasising that some meaning aspects are not predictable from the adjective or the noun, but originate from the NP construction. This is a fully fledged and productive naming mechanism in Dutch, competing with adjective-noun compounds (De Caluwe, 1990; Booij, 2002). Such adjective-noun sequences constitute conventional expressions to designate the referent in examples such as *het stoffelijk-∅ overschot* ('the mortal remains') or *het meewerkend-∅ voorwerp* ('the indirect object'). It is worth mentioning that the Dutch reference dictionary *Van Dale* (Geeraerts, 2003) lists *groot seminarie* as an adjec-

tive-noun compound, *grootseminarie*, illustrating the fuzzy boundaries between both naming mechanisms.

The semantic parameter also covers cases where the uninflected adjective realises a semantic specialisation. Whereas the NP with the inflected adjective *het Nederlands-e elftal* ('the Dutch football team') may refer to any football team with (a lot of) Dutch players, the uninflected counterpart *het Nederlands-∅ elftal* narrows the reference to the Dutch national team (Honselaar, 1980).

Summarising, according to the semantic parameter, the absence of the adjectival inflectional ending is considered to be a formal indication of semantic specialisation. The morphologically marked form, i.e. the uninflected adjective, iconically identifies a marked meaning.

A closer examination of the semantically oriented analyses of the inflectional alternation reveals three methodological peculiarities. First, the analyses invoking the non-compositional semantics of the NP as a factor motivating the use of the uninflected adjective, do not agree upon the obligatoriness of the uninflected adjectival form to identify a NP with non-fully compositional semantics. Whereas Odijk (1992: 202) states that "when the adjective is combined with the noun to form a N' [a NP with lexical semantics], it gets no *-e*", Booij (2002: 316) considers the use of the uninflected adjective rather as a tendency. Second, the examples quoted in the literature are usually highly idiomatic and tend to neglect adjective-noun combinations with transparent semantics. Finally, although the semantic parameter is considered a substantial argument to account for the use of the uninflected adjective, it has not yet been subject of an empirical analysis. Previous empirical studies of the adjectival inflection (Lebrun and Schurman-Swillen, 1966; de Rooij, 1980) did not measure the impact of this parameter, invoking seemingly insurmountable methodological difficulties to make it operational. All "fixed combinations" (de Rooij, 1980: 12) are excluded, without making explicit the criteria to decide when an adjective-noun combination is considered to be "fixed".

These observations motivate a thorough empirical analysis. This is the only way to evaluate the actual impact on the inflectional alternation of the frequently quoted semantic parameter. Furthermore, an empirical analysis would enable us to check whether the uninflected adjective is indeed obligatory in idiomatic expressions. However, the interpretation of the semantic parameter in terms of semantic specialisation or markedness, as is the case in most introspective analyses, is difficult to apply in an empirical analysis, let alone quantify. As a consequence, we will focus on another factor characteristic of idiomaticity, namely the collocational value of the adjective-noun combination. Geeraerts (1986: 134) defines lexical collocation as "idiosyncratic restrictions on the combinatorics of lexical elements". This definition ranges from expressions with an opaque meaning, such as *het hard-∅ gelag* ('the bad break'), to formally fixed expressions with a transparent meaning, such as *het blond(-e) bier* (lit. 'the blond beer'; 'the lager beer'). This approach agrees with lexical analyses, where idiomaticity is not considered as a discrete but rather as a gradient property (Cruse, 1986; Geeraerts, 1986; Van Sterkenburg, 1993).

We do not *a priori* exclude an analysis of the semantic effects mentioned above, but in this contribution we want to evaluate the explanatory power of lexical collocations. In the next section, we will present a quantitative alternative to compute the lexical attraction between the adjective and the noun in a NP by means of the log-likelihood ratio test.

3. Quantifying lexical collocations

The goal of the quantitative test is to compute the lexical attraction strength between the adjective and the noun comparing their expected and observed frequencies. The underlying assumption is that related words will tend to co-occur more often than they would by pure chance. Applied to the present topic, adjectives and nouns constituting a lexical collocation will tend to co-occur more frequently than what would be predicted using their unigram frequencies. Moreover, this approach enables the detection of usage effects, i.e. tendencies in language use structuring the grammar bottom-up (Bybee and Hopper, 2001). In brief, this quantitative approach makes a semantic effect operational by means of a combinatorial criterium based on usage effects.

Before presenting the measure to quantify the lexical attraction strength, it should be noted that lexical (Cruse, 1986) and morphological handbooks (Matthews, 1991) mention qualitative tests to identify lexical collocations. The rationale underlying the qualitative tests is to check whether the adjective-noun bigram possesses phonetic, morphological, syntactic and semantic properties commonly associated with NPs. Examples of such criteria are the absence of the common projection possibilities of the adjective and the noun, and the impossibility to substitute the adjective and the noun by (near-)synonyms. Lexical collocations share more properties with adjective-noun compounds than with common NPs. Although the application of those criteria allows the construction of a scale ranging from a high to a low collocational score, we choose not to apply it because these criteria rest on introspective and subjective judgements. Besides, it is not clear how to establish a hierarchy between the different criteria in order to resolve problems resulting from conflicting criteria (Matthews, 1991). In future research, the impact of these criteria must be accounted for.

Although we realise that it is difficult – if not impossible – to select the procedure that provides the best assessment of collocations[2], the log-likelihood ratio test (henceforth LL; Dunning, 1993) seems to be the best fit for the present purpose compared to other common collocation measures, such as Pearson's χ^2, the mutual information index (Church and Hanks, 1990) and the t-test. First, best results in identifying adjective-noun collocations in a pre-established list are achieved using LL. Second, LL performs well with sparse data (Dunning, 1993; Manning and Schütze, 2002). The other tests mentioned all appear to have some major drawback. The t-test assumes a normal distribution, which cannot be taken for granted in natural language (Manning and Schütze, 2002). As for the χ^2-test, its performance decreases with low frequency data. This objection also applies to the mutual information index, which systematically overestimates low frequencies.

The LL ratio compares the probability that the adjective (ADJ) and the noun (N) are independent (zero hypothesis) to the probability that ADJ and N are dependent (research hypothesis), by verifying whether the presence of N given ADJ is significantly different from N not given ADJ. Applied to *het groot-∅ seminarie*, this strategy amounts to verifying whether *seminarie* is significantly more frequent than any other noun after *groot-∅*, than after any other adjective. When this is the case, the sequence gets a high LL score; in the opposite case, it receives a low score.

[2] We do not use the association measure to identify or to extract potential lexical collocations from a text, but to compute the collocational strength between an adjective and a noun in a pre-established list. Consequently, we do not have to bother about the optimal span for the collocates, since the adjective and the noun are adjacent within the domain of agreement, viz. the NP (see section 2).

	ADJ	**¬ADJ**	**Σ**
N	a	b	a+b
¬N	c	d	c+d
Σ	a+c	b+d	n

Table 1. Contingency table to compute -2LL

Based on the 2x2 contingency table presented in table 1, the LL ratio is defined by means of the following formula: -2LL = 2(a ln(a) + b ln(b) + c ln(c) + d ln(d) + n ln(n) – (a+b) ln(a+b) – (c+d) ln(c+d) – (a+c) ln(a+c) – (b+d) ln(b+d)).

4. Empirical analysis

Before discussing and evaluating the results of the empirical analysis (section 4.2.), we will briefly present the corpus and the way we scrutinised it (section 4.1.).

4.1. *Gathering the data*

The data used in the empirical case-study are extracted from the *Corpus Gesproken Nederlands* ('Corpus Spoken Dutch'; henceforth CGN) (Oostdijk, 2001). This is a 5m words database of spoken Dutch[3]. CGN covers both lectal distinctions that are considered relevant for the alternation between inflected and uninflected adjectives: geographically, Netherlandic and Belgian Dutch data are gathered ; from a stylistic point of view, the corpus comprises different varieties of spoken Dutch, ranging from controlled monologues to spontaneous and informal face-tot-face dialogues and multilogues.

To compute the -2LL of an adjective-noun bigram, the following data have to be gathered: the frequency of the adjective-noun bigram under consideration, the number of bigrams containing ADJ but not N, the number of bigrams containing N but not ADJ, and the number of bigrams containing neither ADJ nor N. To this end, a list is compiled of all bigrams occurring at least three times in CGN (n = 3237915) and their respective token frequencies. All bigrams occurring less than three times in the corpus are excluded, because they would yield statistically insignificant results. Next, this overall bigram list is compared with the database of adjective-noun bigrams appearing in a NP with a singular neuter head noun and a definite determiner[4] (n = 4970), viz. the agreement domain where the inflectional variation under investigation shows up (see section 2). This matching operation yields a list of 1489 adjective-noun bigrams for which a -2LL score is computed. This -2LL value is used to rank the bigrams. The remainder of the database consists of bigrams occurring less than three times in CGN. Those bigrams (n = 3481) are assigned a -2LL value of 0, the lower limit of the -2LL ranking.

4.2. *Evaluating the results*

The output of the LL test will be processed in two stages. First, the -2LL ratio will be plotted to the distribution of inflected and uninflected adjectives in order to verify the empirical

[3] The final release (release 7) of CGN, which is expected in January 2004, will comprise 10m words. The present analysis uses release 5 (April 5, 2002), containing about 5m words.

[4] Adjectives that are always inflected, such as the superlative in attributive position (*het mooist-e paard*/**het mooist-∅ paard* 'the most beautiful horse'), or always uninflected, such as adjectives ending in a vowel (**het lila-e jasje/het lila-∅ jasje* 'the lilac jacket'), are removed from the database, since they do not exhibit the variation under investigation.

validity of lexical collocations as a parameter to interpret the alternation between inflected and uninflected adjectives. Second, the -2LL score will be entered, together with the other potential explanatory parameters mentioned in section 2, as regressor variables in a binomial logistic regression with the probability of the uninflected adjective, or more precisely the odds of the probability of the uninflected adjective, as response variable. This multifactorial statistic will be used to quantify the impact of the semantic parameter and to compare its relative strength to that of the other regressors in the model. At this end, the relevant adjective-noun bigrams (n = 4970) are analysed with respect to the other relevant parameters influencing the alternation between inflected and uninflected adjectives (see section 2)[5].

As mentioned in the previous section, only 1489 out of the 4970 adjective-noun bigrams occur at least three times in CGN. This entails that 3481 observations are assigned a -2LL score of 0. Table 2, where 'abs' stands for absolute frequency and 'rel' for relative frequency, shows that the subset of the observations with a -2LL higher than 0 contains significantly more uninflected adjectives than the subset with a -2LL equal to 0 (df = 1; $\chi^2 = 257.7166$; $p \leq 0.001$)[6].

adjective	-2LL = 0		-2LL > 0		Σ	
	abs	rel	abs	rel	abs	rel
inflected	2891	0.8305	924	0.6206	3815	0.7676
uninflected	590	0.1695	565	0.3794	1155	0.2324
Σ	3481	1.0000	1489	1.0000	4970	1.0000

Table 2. Distribution of both adjectival forms in bigrams with -2LL = 0 and -2LL > 0

These figures indicate a positive effect of highly recurrent bigrams on the use of the uninflected adjective. Focussing on the subset with -2LL > 0, this tendency can be analysed in further detail. First, this subset displays a positive correlation between the -2LL score and the use of the uninflected adjective as demonstrated by the positive correlation coefficients between the cumulative probability of the uninflected adjective at the one hand and the -2LL score and the -2LL ranking at the other:

- $r(Pr_{cum}(ADJ\_uninflected),LL\_score) = 0.48294$
- $r(Pr_{cum}(ADJ\_uninflected),LL\_ranking) = 0.98709$

Second, the 1489 bigrams tokens for which -2LL > 0 represent 345 bigram types. For 289 of those bigram types (83.77%), all occurrences allow one adjectival form. In other words, bigrams where the adjective and the noun co-occur more often than would be expected by pure chance manifest a strong tendency to favour the use of one inflectional alternative, either the inflected or the uninflected, to the detriment of the other. This result puts the generally accepted view that idiomaticity motivates the use of uninflected adjectives in a broader perspective. It will have to be verified whether the use of the inflected adjective in such cases is motivated by semantic or other parameters.

[5] The empirical analysis is performed by means of *Abundantia Verborum*. This is a computer tool to query corpora and to label and classify the extracted data. For more information, we refer to Speelman (1997) or http://wwwling.arts.kuleuven.ac.be/genling/abundant.

[6] The subset with -2LL = 0 shows a very homogeneous distribution of both adjectival alternatives as indicated by the low variance of the cumulative probabilities of the uninflected adjective ($\sigma^2 = 0.000426$).

A closer look at the data reveals that no inflected adjective appears amongst the 196 tokens or 11 types with the highest -2LL values. Of the 50 bigram types (417 tokens) at the top of the -2LL ranking, only three bigram types have a highly idiomatic meaning, but remarkably, the adjective takes an *-e* suffix in all three cases: *het Oud-e Testament* ('The Old Testament'), *het Groen-e boekje* (lit. 'the little green book', hence 'the standard word list of Dutch) and *het groen-e Poldermodel* (lit. 'the green Polder model', hence 'the political model for polder ecology'). The other bigrams in this top list are lexical collocations, such as *het Europees-∅ kampioenschap* ('the championship for European national teams') or *het cultureel-∅ centrum* ('the cultural centre'). Amongst these sequences, we find conventional names for official institutions, such as *het openbaar-∅ vervoer* ('the public transport') or *het secundair-∅ onderwijs* ('secondary education').

Summarising, compared to adjective-noun combinations with a loose association strength, sequences with a strong association between adjective and noun seem to favour the use of the uninflected adjective. Moreover, such bigrams show little internal variation between both inflectional alternatives, revealing the importance of usage effects. On the other hand, the data clearly show that the use of the uninflected adjective is not a binary but a gradient matter. In order to gain insight in the effective impact of the semantic parameter, we will enter it as a regressor variable in a logistic regression together with the factors mentioned in section 2. At this point, it is important to mention that the resulting regression model is not the end of the research. The other parameters need further refinement and extension.

A logistic regression analysis[7] (Rietveld and van Hout, 1993: chap. 9; Paolillo, 2002: chap. 8) allows the use of categorical and numeric variables. Moreover, since an additive model may predict values outside the range [0,1], the dependent variable is transformed in a logit function, compelling the use of a non-linear statistic, viz. the logistic regression. This statistic computes the best fitting logistic curve for a given distribution applying maximum likelihood estimation to determine the coefficients of the regressor variables. We will use the logistic regression not only to quantify the impact of lexical collocations on the inflectional alternation, but also to estimate its relative impact compared to other potential explanatory factors quoted in the literature by comparing their respective estimates. Note that the logistic regression calculates changes in the log odds of the dependent variable, not changes in the dependent variable itself. Therefore, standardised estimates are calculated, the so-called β weights. The β weights of the factor values are interpreted as follows: a β weight equal to 5 for a factor value indicates that the odds ratio uninflected vs. inflected is predicted to increase 5 times as a result of the effect of that factor. Conversely, a β weight equal to 0.2 means that the odds ratio uninflected vs. inflected is predicted to decrease 5 times as a result of the effect of that factor value.

The potential explanatory factors mentioned in section 2 are made operational as follows:

- part of speech of the determiner: definite article (*het* 'the'), demonstrative pronoun, possessive pronoun, genitive determiner
- adjectival length: 1 syllable, 2 syllables, 3 syllables, more than 3 syllables
- accentuation of final adjectival syllable: not accentuated, accentuated
- register: monologue, dialogue/multilogue
- region: Netherlandic Dutch, Belgian Dutch
- lexical collocation: -2LL of the adjective-noun bigram

[7] The regression analysis has been conducted using the R statistical package. More information can be found on http://www.r-project.org/.

Note that only the -2LL factor is numeric. For the categorical factors, the first value is the control value to which the other values are compared in order to determine their effect on the response variable. The variables are added to the model by means of the forward stepwise selection method. This selection procedure starts with an empty model and adds at each stage the best performing regressor variable to the model, namely the regressor variable performing the strongest reduction of the unexplained variation. Table 3 outlines the building of the regression model applying the forward stepwise selection procedure. The unexplained variation is computed by means of AIC[8].

Step	AIC	Regression model
Start	5391	
Step 1	4697.02	response.variable = -2LL.ADJ.N
Step 2	4093.97	response.variable = -2LL.ADJ.N + adj.length
Step 3	3974.37	response.variable = -2LL.ADJ.N + adj.length + register
Step 4	3881.96	response.variable = -2LL.ADJ.N + adj.length + register + adj.final.accent
Step 5	3821.30	response.variable = -2LL.ADJ.N + adj.length + register + adj.final.accent + region
Step 6	3815.19	response.variable = -2LL.ADJ.N + adj.length + register + adj.final.accent + region + determiner

Table 3. Building the regression model using the stepwise forward procedure

Since the -2LL ratio is the first parameter to be entered to the regression model, it can be considered the best explaining factor for the use of the uninflected adjective. The data in table 3, however, need further refinement. Therefore, let's turn to table 4. This table summarises the impact of the different factor values included in the logistic regression analysis.

regressor	estimate	β weight	95% CI β weight	
			Lower	Upper
(intercept)	-3.2139			
det.demonstrative	-0.2392	0.7873	0.5975	1.0373
det.possessive	0.3707	1.4488	1.1223	1.8703
det.genitive	0.1820	1.1996	0.3419	4.2096
adj.length.2syl	0.8056	2.2381	1.6778	2.9855
adj.length.3syl	1.9844	7.2745	5.6037	9.4435
adj.length.>3syl	2.0927	8.1070	5.9808	10.989
adj.final.accent.yes	-0.9282	0.3953	0.3290	0.4749
register.dial/multil	1.0433	2.8386	2.4008	3.3562
region.belgium	0.6718	1.9577	1.6541	2.3170
-2LL.ADJ.N	0.0101	1.0102	1.0089	1.0114

Table 4. Result of the logistic regression analysis

First, we will discuss the effect of the individual factor values in the model. Looking at the confidence intervals, all regressors, except the demonstrative pronoun and the genitive determiner, appear to differ in a significant way from the control value. The effect of these regres-

[8] AIC (Akaike Information Criterion) is an adjusted measure to compute the (un)explained variation in a fitted model. Due to a bias, being a function of the number of degrees of freedom in the model, maximum likelihood estimation is not suited to compare different fitted models. This bias is adjusted by AIC.

sors does not significantly differ from the definite article *het* ('the'). Moreover, the factor values corroborate the tendencies described in section 2. Comparing the impact of the different factors in the regression model, the -2LL ratio is the best performing parameter. It is not only the first parameter to be added to the stepwise regression model as shown in table 3, but it also has the greatest effect on the predictive power of the model. Although its coefficient estimate and β weight are relatively small compared to those of other significant values, it should be kept in mind that this factor is not binary (0 vs. 1), as is the case for the categorical factor values, but numeric, with a mean value of 60.65.

Next, we will assess the quality of the fitted model examining its predictive and explanatory power. When used to predict the adjectival inflection in the 4970 observations in the database, the fitted model achieves a success rate of 83.76%. This score is an important improvement with respect to the intercept only model, which has a success rate of 76.76%. Let's turn to the explanatory power of the fitted model. This can be assessed looking at the variation explained by the fitted model, which is the remainder of the total variation in the intercept only model minus the unexplained variation in the fitted model: 5389.00 – 3793.18 = 1595.82. This figure yields a model χ^2 $p \leq 0.001$ (df model $\chi^2 = 10$). Hence, we may conclude that the fitted model reduces the unexplained variation in a highly significant way.

The logistic regression analysis shows that the semantic parameter is the most important factor to account for the use of the uninflected adjective. Nevertheless, it is not a sufficient factor, as indicated by the significant impact of other parameters and the success rate of the regression analysis which is inferior to 100%. We also have to bear in mind that we used a combinatorial criterium, namely the attraction strength within lexical collocations, to test the semantic effect, in contrast to introspective analyses, which mostly rely on idiomaticity. Furthermore, the effects of the other parameters included in the regression model need to be analysed in further detail.

5. Conclusions

In this paper, we have presented an approach to empirically validate semantic effects in a case study of variational linguistics, namely the alternation between inflected and uninflected Dutch attributive adjectives. The semantic effect of idiomaticity was made operational in terms of lexical collocations, looking for co-occurrence patterns within adjective-noun bigrams. The notion of lexical collocation was argued to seize the insight that idiomaticity is not a binary, but a gradient notion. Furthermore, lexical collocations can be quantified in a straightforward way, computing the lexical attraction between the adjective and the noun.

The empirical analysis revealed a clear influence of lexical collocations on the inflectional variation, favouring the use of the uninflected variant when the adjective and the noun co-occur more frequently than would be predicted by pure chance. Furthermore, tightly associated adjective-noun combinations showed little internal variation between both adjectival alternatives. In order to seize the real influence of lexical collocations on the inflectional alternation, we entered it, together with other relevant factors, in a logistic regression analysis. This analysis corroborated the widespread opinion amongst Dutch linguists that the semantics of the adjective-noun combination is the predominant motivation to use the uninflected adjective.

Different issues remain to be tackled in future research. The quantitative results of the collocation analysis must be completed with a qualitative interpretation, focussing on other, more semantic characteristics of idiomaticity. With respect to the bigram types allowing only one

inflectional variant, elements determining the choice for the given alternative must be looked for. As for the regression analysis, the regressor variables have to be refined and extended in order to increase the explanatory power of the model.

References

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

Cruse D.A. (1986). *Lexical Semantics*. Cambridge University Press.

Booij G. (1992). Congruentie in Nederlandse NP's. *Spektator*, vol. (21/2): 119-135.

Booij G. (2002). Constructional Idioms, Morphology, and the Dutch Lexicon. *Journal of Germanic Linguistics*, vol. (14/4): 301-329.

Bybee J. and Hopper P. (Eds) (2001). *Frequency and the emergence of linguistic structure*. John Benjamins.

Church K.W. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. (16): 22-29.

De Caluwe J. (1990). Complementariteit tussen morfologische en in oorsprong syntactische benoemingsprocédés. In De Caluwe J. (Ed.), *Betekenis en Productiviteit* (*Studia Germanica Gandensia*). *Seminarie voor Duitse taalkunde*: 9-24.

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

Evert S. and Krenn B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, July 6-11: 188-195.

Geeraerts D. (1986). *Woordbetekenis: een overzicht van de lexicale semantiek*. Acco.

Geeraerts D. (Ed.) (2003). *Van Dale Groot woordenboek der Nederlandse taal*. CD-ROM edition, version 1.3. Van Dale Lexicografie.

Honselaar W. (1980). On the semantics of adjective-noun combinations. In Barentsen A.A., Groen B.M. and Sprenger R. (Eds), *Studies in Slavic and General Linguistics*, vol. (1): 187-206.

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

Manning Chr. and Schütze H. (2002). *Foundations of statistical natural language processing*. MIT Press.

Matthews P.H. (1991). *Morphology*. Cambridge University Press.

Odijk J. (1992). Uninflected Adjectives in Dutch. In Bok-Bennema R. and van Hout R. (Eds), *Linguistics in the Netherlands*. John Benjamins: 197-208.

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

Paolillo J.C. (2002). *Analyzing Linguistic Variation. Statistical Models and Methods*. CSLI Publications.

Rietveld T. and van Hout R. (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.

Rooij J. de (1980). Ons bruin(e) paard. *Taal en Tongval*, vol. (32): 3-25: 109-129.

Speelman D. (1997). *Abundantia Verborum. A computer tool for carrying out corpus based linguistic case studies.* PhD dissertation, KU Leuven.

Van Sterkenburg P. (1993). Gelexicaliseerde woordgroepen van het type A+N. *Tabu*, vol. (23/1-2): 131-142.