Regressing on *er*. Statistical analysis of texts and language variation

Stefan Grondelaers, Dirk Speelman, Dirk Geeraerts

KU Leuven – Department of Linguistics – - Blijde-Inkomststraat 21– 3000 Leuven – Belgium

Abstract

Building on an extensive written corpus of formal and informal Belgian and Netherlandic Dutch, this study tackles the complex distribution of non-anaphoric *er* "there" in adjunct-initial presentative sentences such as *Op de hoek van de straat is (er) een winkel* "At the corner of the street (there) is a shop". The Dutch standard grammar ANS maintains that for this distribution "no strict rules can be given. It can be optional, there may be semantic or stylistic differences, and there is a lot of individual, sometimes also regional variation (1997: 473)." In order to test this view, we confronted two language-structural factors – *adjunct type* and *verbal specificity* – and two contextual factors - *region* and *register* – in a regression analysis of the use of *er* in the 1905 adjunct-initial presentative sentences in the corpus. This statistical analysis demonstrates that the ANS is inaccurate and far too pessimistic as far as presentative *er*'s postverbal distribution is concerned. The fact that language-structural and contextual factors are put on a par in the quoted passage inadequately reflects the far greater impact of the structural factors on *er*'s distribution. In addition, the predictive success of the global *er*-model distilled from the data is strikingly at odds with the "no strict rules"-pessimism of the ANS. The most important discovery in this paper, however, is that *er's* distribution in the Belgian and Netherlandic materials is accounted for by proportionally and structurally different models. The practical consequence of this finding is that the ANS should devote separate entries to *er*'s distribution in Belgian and Netherlandic Dutch.

Keywords: text corpora, logistic regression, contextual and language-structural variation, existential sentences

1. Introduction

Empirical studies of variation in language use basically come in two forms: sociolinguistic, dialectological, register-based studies focusing on lectal forms of variation, and discourse-based studies focusing on the structural and pragmatic factors which influence the presence of one or another grammatical phenomenon within a given text. Quite a number of linguistic phenomena, however, exhibit co-variation of these "external" and "internal" causes of variability. In this paper, we will present an example of how such intricate and rarely studied complexes of variation may be disentangled through the statistical analysis of text corpora.

We focus on the distribution of presentative *er* "there" in adjunct-initial presentative sentences of the type illustrated in (1)-(3):

1. a Een paar weken geleden was *er* ook geen maan. (De Aanslag, p. 50)
   A few weeks ago was *er* also no moon
   A few weeks ago there wasn't any moon either
   b *Een paar weken geleden was ook geen maan.

A few weeks ago was also no moon

2. a In Nederland zijn *er* meer symfonie-orkesten. (NRC 16/05/1992)
In the Netherlands are *er* more symphony orchestras
In the Netherlands there are more symphony orchestras

b In Nederland zijn meer symfonie-orkesten.
In the Netherlands are more symphony orchestra

3. a ?Op de plaats van de Orstkommandatur stond *er* een nieuw bankgebouw.
On the place of the Ortskommandatur stood *er* a new bank building
On the place of the Ortskommandatur *there* stood a new bank building

b Op de plaats van de Ortskommandatur stond een nieuw bankgebouw. (DA220)
On the place of the Ortskommandatur stood a new bank building
On the place of the Ortskommandatur stood a new bank building

The distribution of postverbal presentative *er* in adjunct-initial sentences is extremely difficult to determine, because the preference for *er* is rarely absolute. Only in sentences like (1) - with a sentence-initial time adverbial and a form of *to be* as the main verb -, *er* is a must for all the speakers of Dutch; in the other examples, the need for *er* is a matter of more or less. To complicate matters, there appear to be regional tendencies: it has often been observed that Belgian Dutch manifests a greater tolerance towards postverbal *er* than Netherlandic Dutch (cf. De Rooij, 1991; Algemene Nederlandse Spraakkunst, 1997).

According to the 1984- and the 1997-edition of the Algemene Nederlandse Spraakkunst[1], no strict rules can be given for the presence of absence of postverbal *er*: "it can be optional, there may be semantic or stylistic differences involved, and there is a lot of individual and sometimes also regional variation in its use" (1997[2]:473, but vide also 1984[1]:820). On p. 477 of the second edition, specifically in connection with the optionality of postverbal *er* in adjunct-initial sentences, the ANS states: "In the standard language, *er* is more easily deleted in sentences with a fronted locative than in other cases. (...) The sentence-internal preference for *er* differs from case to case; for the time being, clear rules cannot be given. There are rather large individual differences in its use. We can say however that there is an outspoken geographical tendency: whether or not it follows a locative adjunct, *er* is retained more often in Belgium (...) and also, albeit to a lesser degree, in the southern regions of The Netherlands (...)."

Only De Rooij (1991) concentrates on possible factors which determine *er*'s postverbal behaviour. In the third article of a series in which he investigates regional variation in the use of *er*, De Rooij reports on a survey he used to assess Belgian and Netherlandic *er*-preferences in locative adjunct sentences. The survey not only confirms that there are regional differences in the preference for *er*; it also demonstrates that the locative character of the adjunct cannot be the only factor which determines the postverbal presence or absence of *er*. Another factor, De Rooij continues, could well be the semantics of the main verb: sentences with the verbum finitum *to be* are preferably constructed with *er*.

Stil, De Rooij upholds the validity of the ANS-statement that "for the presence or absence of *er*, no strict rules can be given" (1984: 820). Hopefully, the findings discussed in this paper as well as the possible attempts at explanation can serve as startingpoints and working hypotheses for a thorough investigation of this problem complex" (1991: 127).

[1] The *Algemene Nederlandse Spraakkunst* is the standard grammar of Dutch, which appeared for the first time in 1984. The second edition appeared in 1997.

In this paper we will try to live up to this expectation by including in a large-scale quantitative investigation the four factors which are known to have an impact on *er*'s distribution. Our investigation is based on – as De Rooij himself suggests – "an extensive corpus from which *all* sentences with a [locative] adjunct in sentence-initial position are extracted, whether or not they contain *er*" (1991: 116-117). The factors included are the contextual variables *region* (*er* is attested more frequently in Belgium than in The Netherlands) and *register* (*er* is attested more frequently in informal than in formal registers), and the language-structural variables *adjunct type* (*er* is attested more frequently in sentences with a temporal adjunct than in sentences with a locative adjunct) and *verbal specificity* (*er* is attested less frequently in sentences with a specific main verb).

The material in this paper is organised as follows. Section 2 focuses on the structure of the corpus on which the research is based, and section 3 concentrates on how the independent variables *region*, *register*, *adjunct type*, and *verbal specificity* are operationalised in this investigation. Section 4 briefly introduces the logistic regression technique, which will be used in section 5 to address the principal questions that were raised in connection with *er's* distribution. To find out whether the pessimistic attitude towards *er's* distribution in the ANS and De Rooij (1991) is justified, we will first assess the predictive power of the global model which contains the four independent variables and the interactions between them. More important in the framework of this paper is the regional issue: on the basis of the regression data, we can find out whether the *er*-variation in the Belgian sample is explained by the same model as the variation in the Netherlandic sample. In section 6, all the findings are summarised, and a suitable alternative for the current *er*-entry in the ANS is suggested.

2. Materials

The empirical foundation of this investigation is the ConDiv-corpus, an extensive text-database compiled for a related research project[2]. Table 1 contains an overview of the corpus components that were used in the present research.

	informal - editorial control			formal + editorial control
	Internet	Newspapers		
	UseNet	Popular newspapers		Quality newspapers
		Regional	National	
N	(2.287.648)		*De Telegraaf* (1.590.581)	*NRC Handelsblad* (1.520.064)
B	(2.449.193)	*Het Belang van Limburg* (1.561.362)	*Het Laatste Nieuws* (1.345.367)	*De Standaard* (1.665.144)

Table 1. Overview of the corpus components and their size in number of tokens

[2] The ConDiv corpus was originally compiled for the research project "Convergence and divergence in the Dutch lexicon", a wide-scale investigation into the diachronic and synchronic status of Belgian and Netherlandic Dutch, funded by VNC-grant 205-41.073. More information about the corpus can be found in Grondelaers e.a. (2000).

The ConDiv-corpus basically consists of two types of attested language use. In the newspaper component a distinction is made between quality newspapers such as *De Standaard* and *NRC Handelsblad* and popular newspapers such as *De Telegraaf* and *Het Laatste Nieuws*. There are national popular papers - *Het Laatste Nieuws* en *De Telegraaf* –, but also *regional* popular papers such as *Het Belang van Limburg* or *De Gazet van Antwerpen.*

In addition to newspaper language, the ConDiv-corpus also contains more informal language data attested on the Internet. From Geeraerts, Grondelaers & Speelman (1999), we know that it is incorrect to conceive of Belgian Dutch as a monostratal language. In-between standard language and dialects, there is at least one intermediate level on which a higher degree of informality coincides with geographical specialisation: the more informal the communication setting, the more regional the Dutch sounds in which (especially Belgian) speakers express themselves. In recent publications this intermediate register is dubbed "tussentaal" (Taeldeman 1992: 33-52), "verkavelings-Vlaams" (Van Istendael 1993: 116), of "soap-Vlaams" genoemd (Geeraerts 1999: 232).

Since it is impossible to determine a priori how many intermediate strata must be distinguished in Dutch, the compilers of the ConDiv-corpus did not, in the first place, look for language use which represents a certain stratum. Instead, we considered the different communicative situations in which written Dutch is produced as independent variables, and the register spoken in these situations as the dependent variable. Hence, the stylistic-stratificational variation in Dutch is accommodated in this study by comparing language data from four different communicative situations, which can be positioned on a stylistic scale which has "informal" and "very formal" as its poles. The three types of newspaper materials – regional popular newspapers, national popular newspapers and quality papers – occupy the highest positions on the formality scale.

The lowest position on the scale – informal Dutch – is represented in this study by language data attested on UseNet, an Internet forum on which surfers debate in "newsgroups", by means of e-mail messages they add to an ongoing discussion. Since e-mail is offline – so that users can reread their contributions before adding them to a "thread" – and since academic Internet operators only tolerate (relatively) serious newsgroups on their net –, the UseNet register is not as informal as Internet Relay Chat, a module in which anonymous users debate online.

Our investigation of *er*'s postverbal distribution is not based on all the components of the ConDiv-corpus. Although the corpus contains large portions of Internet Relay Chat, no IRC-material was included in the analysis. The typical interactional characteristics of IRC – "temporality and immediacy" (Bays 1998) – necessitate specific formulation techniques "to augment the speed and the capacity of information transfer" (idem). The most important strategies in this respect are "abbreviation, ellipsis and a telegraphic style, which reduce the quantity of words that need to be typed, sent and read" (idem). Needless to say that postverbal *er* is an endangered linguistic species in the context of this condensed style, especially there were it is not needed for grammaticality[3]. As a result, we restrict the analysis to language materials in which production speed plays no role.

Production speed, however, is not the only menace to a representative *er*-distribution. Sceptics like Verkuyl (1998: 63 ff.) also mention "proof reader-idiosyncrasies": it goes without saying

[3] Grondelaers (2000:193-196) demonstrates beyond doubt that the condensed and elliptical style of IRC has a detrimental effect on the use of *er*.

that postverbal *er* is particularly vulnerable to the last minute interferences of press revisers or editors-in-chief. Some papers, in addition, are known to have style guides which are excessively hostile to the "unnecessary" postverbal use of *er*. To reduce the impact of such hostility, we have restricted the analysis to language data from newspapers which, when asked about their attitude towards *er*, emphatically stated not to have any such policy. None of the Belgian newspapers turned out to devote special attention to *er*; *De Telegraaf* and *NRC* were the only Dutch newspapers which responded to our question, and they too assured us that *er* is in no way stigmatised in their publications.

	UseNet	Popular newspapers		Quality papers
N	n = 192	*Telegraaf* n = 227		*NRC* n = 263
B	n = 225	*Het Belang van Limburg* n = 397	*Het Laatste Nieuws* n = 198	*De Standaard* n = 403

Table 2. Distribution of observations over the sources in the corpus

From the reduced ConDiv-corpus, we subsequently extracted all the adjunct-initial presentative sentences with and without *er*[4]. The distribution of the extracted observations over the different sources in the corpus is given in table 2.

3. Independent variables

The contextual variables *region* – Belgian vs. Netherlandic Dutch – and *register* – UseNet vs. popular newspapers vs. quality papers – are reflected in the structure of the corpus. The extracted observations were tagged for the language-structural variables *adjunct type* and *verbal specificity*.

Adjunct type was operationalised straightforwardly by contrasting observations with a locative and a temporal adjunct[5]. The *verbal specificity* factor, by contrast, does not translate easily into an operational parameter. Although adjunct-initial presentative sentences allow only a limited number of verbal classes, the specificity of those verbs may be determined by any of the three conceptual ingredients their semantics presuppose: nearly all verbs in adjunct-initial presentative sentences such as *On the roof was a bird* code a relation between the referent of the subject (*a bird*) and the temporal or locative setting the adjunct refers to (*the roof*). Because of the conceptual inseparability of a verbal process from its subject and its setting, we have operationalised the *verbal specificity*-factor on the basis of the *size* of the class of

[4] For the extraction, *Abundantia Verborum* was used, a powerful computertool developed by the second author to query corpora, label and classify the extracted data, and perform statistical analyses on them. More information about *Abundantia Verborum* can be found in Speelman (1997), but also on http://wwwling.arts.kuleuven.ac.be/genling/abundant.

[5] Observations like *Bij die ramp vielen 34 doden (hbvl/5400)* "In this disaster 34 people were killed" demonstrate that the opposition locative vs. temporal is not a binary distinction, because *bij die ramp* allows at the same time a locative (on the place of the disaster) and a temporal interpretation (at the time of the disaster). Because of its low frequency, observations with this intermediary adjunct type are excluded from the analysis.

possible subjects the different verbs in adjunct-initial presentatives subcategorise. Building on this criterion, the lowest level of specificity is represented by the verb *to be*, which imposes no restrictions at all on process, setting and subject. In this respect it is hardly surprising that almost all the verbs we encounter in adjunct-initial presentative sentences are hyponyms of the verb *to be*.

The highest level of specificity, by contrast, is represented by verbs which are constructed with a limited set of subjects. These include Levin's (1993: 250) VERBS OF EXISTENCE, which are "typical of certain entities". The latter can be subdivided in VERBS OF ENTITY-SPECIFIC MODES OF BEING (*vloeien* "flow", *branden* "burn"), VERBS OF MODES OF BEING INVOLVING MOTION (1993: 251) like *wapperen* "flutter", and MEANDER VERBS like *meanderen* "meander". VERBS OF SOUND EXISTENCE (1993: 252) like *echoën* "echo" restrict their subject to sound-producing entities, whereas VERBS OF GROUP EXISTENCE (1993: 253) like *zwemmen* "swim" or *dansen* "dance" typically refer to the existence of resp. fish, bees and ants. On the highest specificity level we also find another subcategory of the VERBS OF EXISTENCE, i.e. Levin's (1993: 255) VERBS OF SPATIAL CONFIGURATION which, since they designate "the spatial configuration of an entity with respect to some location", impose specific restrictions on relation and location and, hence, limit the class of potential subjects. Typical examples are *sit*, *stand*, *lie*, and *hang*. The same goes for a second subclass of verbs on the highest specificity level, the VERBS OF APPEARANCE, DISAPPEARANCE AND OCCURRENCE (1993: 258-261), which refer, respectively, to "the appearance of an entity on the scene" (cf. *verschijnen* "appear", *landen* "land", etc.), "the disappearance or going out of existence of some entity" (cf. *perish* "vergaan")[6], and "the occurrence of an event" (*plaatsvinden* "occur", *aan de gang zijn* "take place"). Finally, two subcategories of VERBS OF MANNER OF MOTION appear in adjunct-initial presentative sentences, dubbed the ROLL- en RUN-VERBS (1993: 264-267), although according to Levin they might just as well be considered as VERBS OF EXISTENCE. *Drijven* "float", *glijden* "glide" and *rollen* "roll" are typical ROLL-verbs, whereas *rennen* "run" and *springen* "jump" are frequently attested examples of RUN-verbs.

Contrary to Van Es & Van Caspel (1971), De Rooij (1991), and Grondelaers & Brysbaert (1996), we do not restrict ourselves to a binary specificity opposition. We add an intermediate level represented by a small group of (frequently attested) verbs, which impose a minimal restriction on one of their conceptual ingredients. The verb *bestaan* "exist" is slightly more specific than *zijn* "to be" because it situates its subject within the metaphysical boundaries of this world. *Ontstaan* "to come into being" adds an inchoative aspect to *zijn*, *blijven* "remain" an imperfective aspect, and *voorbij gaan* "to pass" and *eindigen* "to end" a perfective aspect. The frequently attested verb *heersen* "to prevail" bestows a greater agentivity on its animate as well as inanimate subjects. So, what all these verbs have in common in addition to their somewhat schematic meaning, is the fact that they impose minimal restrictions on their subjects, without, however, becoming as unrestricted as *to be*.

4. Logistic regression

Table 3 shows absolute and relative frequencies of *er* as a function of the four factors investigated in this study. In the top row, the Netherlandic (N) and the Belgian sample (B) are contrasted, whereas on the lower row, we distinguish between Use(net), Pop(ular

[6] English does not allow presentative sentences with DISAPPEARANCE-verbs (cf. Levin 1993: 260), but Dutch does: *Op de Noordzee verging gisterenavond een Maltese tanker* *"On the North sea perished yesterday evening a Maltese tanker".

newspapers), and Qua(lity newspapers). The outer vertical axis contrasts presentative sentences with tem(poral) and loc(ative) adjuncts, and the inner vertical axis distinguishes between the three verb types, *zijn* vs. int(ermediary verbs) vs. verbs of EXISTENCE & APPEARANCE.

			N						B						***avg. +er***
			Use		Pop		Qua		Use		Pop		Qua		
			abs	rel	abs	rel	abs	rel	abs	rel	abs	rel	abs	rel	
tem	*zijn*	*-er*	0	0,0	0	0,0	1	2,3	0	0,0	0	0,0	0	0,0	
		+er	35	**100,0**	32	**100,0**	43	**97,7**	40	**100,0**	111	**100,0**	48	**100,0**	***99,7***
	int.	*-er*	3	13,6	11	22,4	14	33,3	1	5,0	16	21,9	18	35,29	
		+er	19	**86,4**	38	**77,6**	28	**66,7**	19	**95,0**	57	**78,1**	33	**64,7**	***75,5***
	e&a	*-er*	0	0,0	8	40,0	4	40,0	8	38,1	20	45,5	13	52,0	
		+er	5	**100,0**	12	**60,0**	6	**60,0**	13	**61,9**	24	**54,5**	12	**48,0**	***57,6***
loc	*zijn*	*-er*	13	33,3	7	36,8	9	37,5	4	7,7	10	10,6	6	15,8	
		+er	26	**66,7**	12	**63,2**	15	**62,5**	48	**92,3**	84	**89,4**	32	**84,2**	***81,6***
	int.	*-er*	47	87,0	48	80,0	72	91,1	29	63,0	95	74,8	101	81,5	
		+er	7	**13,0**	12	**20,0**	7	**8,9**	17	**37,0**	32	**25,2**	23	**18,5**	***20***
	e&a	*-er*	36	97,3	47	100,0	63	98,4	34	73,9	137	93,8	111	94,9	
		+er	1	**2,7**	0	**0,0**	1	**1,6**	12	**26,1**	9	**6,2**	6	**5,1**	***6,3***
avg + er				***48,4***		***46,7***		***38***		***66,2***		***53,3***		***38,2***	

Table 3. Absolute and relative frequencies of er *in adjunct-initial presentative sentences as a function of region, register, adjunct type and verbal specificity*

The data in table 3 were collected in order to answer four questions in connection with postverbal *er*. First, is the impact of the individual variables on *er*'s distribution statistically significant? Second, which factor's impact is the most outspoken? Third, is the explanatory and predictive power of the model which contains these variables as poor as traditional analyses – notably the ANS and De Rooij (1991) – would like us to believe? Fourth and foremost, can the *er*-variation in the Belgian and Netherlandic samples be explained and predicted by the same model?

For a statistically sound answer to these questions, the data are subjected to a logistic regression analysis. The model equation

4. logit ER = $a + x$b + yc + zb*c

contains, next to the constant *a*, the independent variables b, c and the interaction b*c, as well as the estimates *x*, *y*, *z* in which the statistical tool SAS expresses the mathematical importance of the independent variables. These estimates are subsequently translated into Odds Ratios that can be straightforwardly interpreted. The statistical significance of the effect of an independent variable is expressed in a p-value, as well as the *confidence interval* of the Odds Ratio, which delimits the interval in which the Odds Ratio finds itself, given the variance in the data and the significance level. Odds Ratios can be used to measure hierarchical relations between the independent variables within the same model. In order to assess the impact of the same independent variable in different models, we compare the confidence intervals of the Odds Ratios for that variable in the models compared.

Odds Ratios are interpreted as follows: if SAS returns Odd Ratio "6" for an independent variable, the use of *er* vs. the non-use of *er* is predicted to increase 6 times as a result of the impact of that independent variable. Conversely, Odds Ratio "0.2" indicates that the use of *er* vs. the non-use of *er* is predicted to decrease 5 times as a result of the impact of that variable.

The SAS-output also contains quantities that can be used to evaluate the global quality of models of independent variables. The *Akaike Information Criterion* (AIC) returns two quantities. The first is a log likelihood ratio which expresses the total amount of variation in the *intercept only-model*, the null-hypothesis model without the independent variables invoked to explain *er*'s postverbal distribution. The second quantity expresses the amount of variation which is left unexplained by the *intercept and covariates-model*, the model which does contain the independent variables. The smaller the second quantity with respect to the first, the more powerful the model.

In addition, SAS returns quantities which reflect the predictive power of a model. In order to determine these measures, SAS considers all the mathematically possible couples of observations whereby the first observation contains *er* and the second observation does not contain *er*, i.e. couples with the structure {+*er*,-*er*}. *Concordant* (C) are those couples whereby the model predicts a higher probability of *er* for the +er-observation than for the –*er*-observation. Discordant (D) are those couples whereby the model predicts a lower probability of *er* for the +*er*-observation than for the –*er*-observation. The g*amma index* – the standardised measure for the relation between concordant and discordant which is used in this study – is calculated on the basis of the formula (C-D)/(C+D). It goes without saying that models with a high gamma-index have a high predictive power.

5. Results and evaluation

Let us start with an analysis of the *er*-variation in the global database. This is the formula for the most powerful model with separate variables and interactions:

5. logit ER = -3.8992 + 3.6941 *adjunct type* + 4.2317 *verbal specificity 1* + 1.1831 *verbal specificity 2* + 1.2121 *region* + 0.9873 *register 1* + 0.4407 *register 2* – 1.2791 *adjunct type*region*

Observe that this model is obtained by regarding *verbal specificity* and *register* as nominal instead of ordinal variables (which would be theoretically possible). A practical consequence is that SAS returns two estimates for these variables, a first one for the impact on *er*'s distribution of the opposition between values "1" and "3" (in the case of *verbal specificity*, between main verb *to be* and a main verb of the EXISTENCE & APPEARANCE-type), and a second for the impact on *er*'s distribution of the opposition between values "2" en "3" (the difference between intermediary verbs and EXISTENCE & APPEARANCE-verbs).

Table 4, which contains the p-values and the Odds Ratios for the independent variables, shows that all the variables and the interactions in the model are highly significant ($p < 0.0001$, except for *register 2* $p = 0.0051$). For the language-structural variables *adjunct type* and *verbal specificity*, whose impact can be readily discerned in table 3, this outspoken statistical significance could be predicted. Slightly more surprising, however, is the high significance of both *register*-variables, whose effect in table 3 is mainly restricted to the Belgian sample. And although regional variation is largely limited to the popular newspapers and, in particular, UseNet, the high significance of the variable *region* from now on necessitates extreme caution when interpreting the supraregional data, especially because *region* also interacts significantly with *adjunct type* in this model.

	p	O.R.
adjunct type	0,0001	40,2
verbal specificity 1	0,0001	68,8
verbal specificity 2	0,0001	3,27
region	0,0001	3,36
register 1	0,0001	2,68
register 2	0,0051	1,55
*adjunct type*region*	0,0001	0,28

Table 4. p-values and Odds Ratios of the independent variables which explain er's distribution in the global database

Despite their outspoken significance, the respective impact of these variables on *er*'s distribution differs noticeably. The fact that *adjunct type* receives an Odds Ratio of 40.2 in this model signifies that the use of *er* compared with the non-use of *er* is predicted to be more than 40 times higher in temporal adjunct sentences than in locative adjunct-sentences. The Odds Ratio 68.8 for *verbal specificity 1* indicates that the use of *er* vs. the non-use of *er* is predicted to increase more than 68 times when an EXISTENCE & APPEARANCE-verb is substituted with a form of *to be*. According to the Odds Ratio 3.27 for *verbal specificity 2*, the effect of the substitution of an EXISTENCE & APPEARANCE-verb with an intermediary verb is less far-reaching. The Odds Ratios for *verbal specificity* indicate, in other words, that the verb *to be* is the main *er*-trigger in an adjunct-initial sentence: postverbal *er* is restricted most efficiently by using a more specific main verb than *to be*; the exact nature of the more specific verb (intermediary or EXISTENCE & APPEARANCE) is not so important according to our data.

The Odds Ratios demonstrate that the impact of the contextual variables on *er*'s behaviour is relatively limited compared to the effect of the language-structural factors. The Odds Ratio 3.36 for *region* reveals that the use of *er* vs. the non-use of *er* is predicted to be more than 3 times higher in Belgium than in The Netherlands. The *register*-variables – resp. Odds Ratios 2.68 and 1.55 – have the smallest impact on *er's* distribution.

The precise impact of de interaction *adjunct type* region*, finally, is difficult to assess, because the decrease in the use of *er* which is suggested by the Odds Ratio 0.28, can be interpreted in two ways: either the regional variation in the use of *er* is restricted to locative adjunct-sentences, or the impact of *adjunct type* is most outspoken in the Netherlandic sample.

Next, let us turn to the quality of the global model. The predictive power of the latter, reflected in the high gamma-index 84.8 %, contrasts sharply with the pessimistic attitude towards *er*'s distribution propounded in the ANS (1984: 820; 1997: 477) and in De Rooij (1991: 127). In the absolute majority of adjunct-initial presentative sentences, *er*'s distribution can be predicted with the simplest of algorithms: neither the use of *er* in a temporal adjunct-sentence with the main verb *to be*, nor the absence of *er* in a locative adjunct-sentence with a more specific verb will ever lead to unacceptable sentences. Judging from the noticeably lower *er*-proportions in the bottom rows of the left half of table 3, the latter goes especially for *er*'s distribution in Netherlandic Dutch.

The latter observation brings us to the question which started off this paper: can we explain the *er*-variation in the Belgian and the Dutch sample by means of the same model? A first

indication can be found in table 5, which contains rudimentary Netherlandic (N) and Belgian (B) models without interactions. The first column of the Netherlandic as well as the Belgian part of table 5 lists the p-values of the independent variables, the second column contains their Odds Ratios, and columns 3 and 4 respectively contain the lower and the upper limits of the confidence interval of the Odds Ratios:

	N				B			
	p	O.R.	confidence intervals		p	O.R.	confidence intervals	
adjunct type	.0001	31,667	18,703	53,615	.0001	11,871	8,267	17,045
verbal spec. 1	.0001	47,177	22,41	99,318	.0001	91,937	52,34	161,493
verbal spec. 2	.0002	3,204	1,718	5,976	.0001	3,299	2,273	4,788
register 1	.0389	1,869	1,033	3,382	.0001	3,418	2,107	5,545
register 2	.0780	1,65	0,945	2,881	.0303	1,515	1,04	2,206

Table 5. p-values, Odds Ratios and confidence intervals of the independent variables which explain er's distribution in the Netherlandic and the Belgian samples

A comparison of the Odds Ratios in the rudimentary Netherlandic and Belgian models instantly reveals two important differences. To begin with, *register 2* is significant in the Belgian model (p = .0303) but not in the Netherlandic model (p = 0.78), reflecting the limited impact of register variation on Netherlandic *er*-preferences. Although, more importantly, *adjunct type* and *verbal specificity 1* continue to be the principal determiners of *er*'s behaviour in the Netherlandic and the Belgian model, their Odds Ratios show that they play a different role in both samples: in the Belgian materials, *adjunct type* and *verbal specificity 1* have a similar impact on *er*'s distribution (O.R. 31.667 and 47.177 resp.), but in the Netherlandic database, the effect of *adjunct type* is distinctly more limited than that of *verbal specificity 1* (O.R. 11.871 and 91.937 resp.).

A more reliable confirmation of the different status of *adjunct type* in the Netherlandic and the Belgian sample can be obtained by comparing the confidence intervals of the Netherlandic and Belgian Odds Ratios for that variable: if it turns out that these intervals do not overlap, we have statistical evidence that there is at least a proportional difference between both models. Table 5 shows that the confidence interval of the Odds Ratio for *adjunct type* in the Netherlandic sample (18.703 - 53.615) does indeed fail to overlap with the confidence interval of the Odds Ratio for *adjunct type* in the Belgian sample (8.267 – 17.045).

The difference between the Netherlandic and the Belgian model becomes even more obvious when we instruct SAS to look for the strongest possible model for the Netherlandic distribution (in 6) and the Belgian distribution (in 7):

6. logit ER = -4.2904 + 4.5131 *adjunct type* + 4.8934 *verbal specificity 1* + 2.4305 *verbal specificity 2* – 0.8474 *adjunct type*verbal specificity 1* – 1.9544 *adjunct type*verbal specificity 2* + 1.5235 *adjunct type*register 1* + 0.438 *adjunct type*register 2*

7. logit ER = -2.7857 + 2.4741 *adjunct type* + 4.5211 *verbal specificity 1* + 1.1935 *verbal specificity 2* + 1.2291 *register 1* + 0.4154 *register 2*

Observe, first, that the rudimentary Belgian model in table 5 is the most powerful model SAS can come up with (both have gamma-index 85.1 %). On the Dutch side, matters are more

complicated: addition of the interactions *adjunct type*verbal specificity 1 & 2* and *adjunct type*register 1 & 2* to the basic model in table 5 results in a better AIC-score (*Intercept + covariates* 472.839 < 477.64), as well as a better gamma-index (89.5 % > 85.2 %), which indicates that the extended model explains and predicts more variation than the rudimentary model. Unfortunately, the added interactions are not all significant: removal of the non-significant interactions *adjunct type*wwspec 1* (p = 0.5177) and *adjunct type*register 2* (p=0.2685), however, causes a statistically unacceptable model.

More importantly, the comparison of the extended models in (6) and (7) confirms what was already suggested by the data in table 5: only the Belgian model retains the *register*-factors as main effects. In the Netherlandic model, the impact of *register* is constrained by *adjunct type*.

These results indicate that there is not only a proportional difference between the Netherlandic and the Belgian model for *er*'s distribution, but also a structural difference, to the extent that the Netherlandic and the Belgian model are made up of different ingredients. Hence, *er* is used not only less in Netherlandic Dutch than in Belgian Dutch, but it is also used differently.

6. Conclusions and consequences

Let us summarise. In this study we have used an extensive corpus of non-elicited contemporary written Dutch to investigate four factors which, according to the linguistic literature, determine the postverbal distribution of *er* "there" in adjunct-initial presentative sentences. The logistic regression we carried out on these factors yields exciting linguistic and methodological consequences.

As far as the linguistic description of *er* is concerned, the regression analysis revealed that the contextual factors *region* and *register* as well as the language-structural factors *adjunct type* and *verbal specificity* have a statistically significant impact on *er*'s distribution. In addition, the data suggest that there is a significant interaction between the factors *adjunct type* and *region*: the difference between the Netherlandic and the Belgian use of *er* appears to be mainly restricted to locative adjunct sentences.

However, the effect on *er*'s distribution of these factors is not identical: whereas the language-structural factors have a massive impact on the use of *er*, the influence of the contextual factors is less outspoken. Yet, the statistical significance of the proportional and structural differences between the Netherlandic and the Belgian use of *er* forces us to formulate separate *er*-models for both regions. The findings at our disposal demonstrate that the *adjunct* factor and the *verbal* factor are absolutely dominant in the Netherlandic model: Netherlandic presentative sentences can usually do without *er* when they have an initial locative adjunct; the verbal factor blocks *er*-preferences in the rare cases where the locative adjunct hasn't already done so. Also, the *er*-determining power of the locative adjunct and the specific verb is not constrained by *register*-variation in the Netherlandic model: the explanatory success of adjunct and verb is more or less identical in all the Netherlandic source types. The result of all this is that *er* is almost absent in Netherlandic locative adjunct sentences with an EXISTENCE & APPEARANCE-verb (n=2).

In the Belgian model, by contrast, the effect of *adjunct type* and *verbal specificity* on *er*'s predictability is less outspoken, and the factors barely interact. In combination with *register*-variation – which does play a role in the Belgian material – this limited predictability leaves substantial *er*-residues (e.g. 26.1 % in UseNet) in contexts which were shown to inhibit the use of *er*, i.e. locative adjunct sentences with an EXISTENCE & APPEARANCE verb.

Methodologically, we hope to have shown that further research into the distributional behaviour of *er* preferably makes use of statistically analysed non-elicited language data. Our regression data have demonstrated in any case that the Belgian and Netherlandic distribution of *er* can be far better explained and predicted than the introspection- and questionnaire-based assertions in the ANS (1984: 820; 1997: 477) and De Rooij (1991: 127) suggest: that "strict rules cannot be given for the time being" is not borne out by our data.

References

Bays, H. (1998). Framing and face in internet exchanges: a socio-cognitive approach. Online publication on http://viadrina.euv-frankfurt-o.de/~wjournal/bays.html

De Rooij, J. (1991). Regionale variatie in het gebruik van *er* III. *Taal en Tongval* 43: 113-136.

Es, G.A. van & P.P.J. van Caspel (1971). De patronen van de zinspotente groepen; grondtype A en zijn varianten II. Publicaties van het archief voor de Nederlandse syntaxis. Groningen: Rijksuniversiteit.

Geeraerts, D. (1999). Noch standaard, noch dialect. 'Tussentaal' in Vlaanderen en Nederland. *Onze Taal* 68: 232-235.

Geeraerts, D., S. Grondelaers & D. Speelman (1999). *Convergentie en divergentie in de Nederlandse woordenschat*. *Een onderzoek naar kleding- en voetbalnamen.* Amsterdam: Meertensinstituut.

Geerts, G., W. Haeseryn, J. de Rooij & M.C. van den Toorn (1984). *Algemene Nederlandse Spraakkunst*. Groningen: Wolters-Noordhoff.

Grondelaers, S. (2000). *De distributie van niet-anaforisch* er *buiten de eerste zinsplaats. Sociolexicologische, functionele en psycholinguïstische aspecten van* er*'s status als* presentatief *signaal*. Doctoral dissertation KU Leuven.

Grondelaers, S. & Marc Brysbaert (1996). De distributie van het presentatieve *er* buiten de eerste zinsplaats. *Nederlandse Taalkunde* 1: 280-305.

Grondelaers, S., K. Deygers, H. Van Aken, V. Van den Heede & D. Speelman (2000). Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde* 5: 356-363.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997[2]). *Algemene Nederlandse Spraakkunst*. Groningen & Deurne: Martinus Nijhoff – Wolters Plantyn.

Istendael, G. Van (1993[10]). *Het Belgisch labyrinth. Wakker worden in een ander land.* Amsterdam: De Arbeiderspers.

Levin, B. (1993). *English verb classes and alternations. A preliminary investigation.* Chicago & London: The Chicago University Press.

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

Taeldeman, J. (1992). "Welk Nederlands voor de Vlamingen?" *Nederlands van Nu* 40: 33-52.

Verkuyl, H.J. (1998). "O corpora, O mores." *Nederlandse Taalkunde* 3: 60-63.