Dutch Strong and Weak Pronouns as a Stylistic Marker of Literariness

Andreas van Cranenburgh

Abstract Certain languages exhibit distinctions between strong and weak forms of pronouns. Linguists have attempted to explain the preferences for the different forms of pronouns in terms of pragmatic factors, specifically discourse salience and contrast. These factors only partially account for the variation observed. In this article we propose to add another factor, style. We investigate the case of Dutch with a corpus of literary novels. We present quantitative results in the form of corpus frequencies and correlations with literary prestige, as well as qualitative judgments from a manual analysis, and finally a statistical analysis of coreference annotations. This complements the linguistic studies, which have focused on testing explanations in specific contexts in controlled experiments, without testing the relevance of those explanations in naturalistic data. Our results suggest that style is a prominent factor in the strong/weak pronoun distinction, since the linguistic explanations have limited predictive power, while our corpus study shows that a high proportion of strong pronouns is associated with literary prestige and Dutch authorship.

Keywords strong and weak pronouns, literariness, Dutch

1. Introduction

What makes a literary novel *literary*? This is a question without an empirically satisfying answer. Various explanations have been suggested. Adherents of Bourdieu claim that the cultural capital of critics and publishers determines the perceived prestige of novels. Proponents of Kantian aesthetics contend that the demarcation of art rests on (inter)subjective, normative value-judgments (i.e., we expect others to agree about the greatness of art, as opposed to matters of taste which are purely subjective). A third group, the formalists, make an even bolder claim, namely that *literariness* is an intrinsic, objective property of texts; proposed mechanisms are defamiliarization and estrangement. Literary language contrasts itself with everyday language by standing out.

van Cranenburgh, Andreas: Dutch Strong and Weak Pronouns as a Stylistic Marker of Literariness, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 217–234. DOI: https://doi.org/10.17885/heiup.1157.c19373

While this paper makes no commitment to any of these explanations, our work most closely aligns with the last explanation, since we will compare objective textual features of texts and correlate them with perceptions of literary prestige. This work is part of a larger research project, The Riddle of Literary Quality,[1] which set out to investigate textual features that may be correlated with literary prestige. To this end, a large reader survey was held (Koolen et al., 2020). Readers from the general public rated 401 recent, best-selling Dutch-language novels (both original and translated) on a Likert scale of 1–7 (not at all literary to very literary). This allows us to estimate the relation between perceptions of literariness and stylistic markers in the texts. Previous work has already shown that literary prestige can be predicted from textual features to a large extent (van Cranenburgh and Bod 2017; van Cranenburgh et al. 2019). The present paper is not about improving on these predictive models, but zooms in on one specific linguistic aspect (the strong/weak pronoun distinction) which turns out to have a surprising correlation with literary prestige (van Cranenburgh et al. 2019), with the aim of better understanding this specific stylistic aspect; i.e., we focus on explanation, not prediction (Breiman 2001).

The Dutch language (along with other languages) has full and reduced versions of some of its personal pronouns (see Table 1). Full pronouns such as *jij* ('you') are also called emphatic or strong, while in Dutch the reduced pronouns such as *je* ('you') are weak pronouns; other types of reduced pronouns such as clitics in Romance languages are grammatically more restricted. On the one hand the distinction follows linguistic rules and cues related to contrast and salience of discourse referents (Bresnan 1998; Kaiser 2011). On the other hand the distinction can also be a stylistic choice, when both options are available. Weak pronouns are more informal and are required in fixed expressions such as *dank je* ('thank you'), whereas strong pronouns can be used for emphasis or to refer to a less salient referent; strong pronouns are required when expressing contrast or in comparisons such as *hij en zij* ('he and she').

This paper addresses the following research questions:

1. Can we explain the large proportion of strong pronouns in some highly literary novels?
2. To what extent is the pronoun form due to a stylistic choice rather than a grammatical preference or requirement?

The rest of this paper is structured as follows. Section 2 goes into the linguistic background of the strong/weak pronoun distinction. Our main results consist of a corpus study (section 3), a study of contrast and preference (section 4), and a statistical model based on coreference annotations (section 5). We end with a discussion of theoretical implications (section 6).

1 https://literaryquality.huygens.knaw.nl.

2. Linguistic Background on Dutch Strong/Weak Pronouns

We first introduce the pronoun system of Dutch and enumerate contexts in which either form is required or preferred. We continue by discussing linguistic theories put forward to explain the choice of pronouns.

2.1 Strong/Weak Pronouns as Described by Reference Grammars

See Table 1 for an overview of Dutch personal pronouns. Some Dutch pronouns have strong and weak forms. These pronouns carry the same meaning, but either the strong or weak variant may be obligatory or preferred in certain contexts. At other times, it is a matter of free choice, i.e., a matter of style.

Table 1 Personal pronouns in Dutch. Pronouns with a common strong and weak counterpart are shown in italics; a comma indicates a subject/object distinction; the forms in parentheses are not common in written language

	Strong	Weak
1st sg	ik, *mij*	-, *me*
2nd sg	*jij*, *jou*	*je*
3rd sg fem	*zij*, haar	ze, (d'r)
3rd sg masc	hij, hem	(ie, 'm)
3rd sg neut	het	('t)
1st pl	*wij*, ons	*we*, -
2nd pl	jullie	-
3rd pl	*zij*, *hen/hun*	*ze*

The Dutch grammar Haeseryn et al. (1997) describes a range of properties of strong and weak pronouns. The most important feature is that phonologically, strong pronouns are often stressed (e.g., when used for emphasis or to contrast a referent with another referent), but weak pronouns are always unstressed. Strong pronouns tend be restricted for persons or concepts treated as persons, while weak pronouns readily refer to both persons and objects. The grammatical contexts where strong pronouns are obligatory are as follows (* marks an ungrammatical phrase; examples adapted from Haeseryn et al. 1997, 252–55):

(1) Comparisons:
ik ben rijker dan *jij*, *dan *je*
'I am richer than *you* (strong), *than *you* (weak)'

(2) Conjunctions of pronouns:
hij en *zij*, *hij en *ze*
'he and *she* (strong), *he and *she* (weak)'

(3) Certain oblique arguments (i.e., neither subject nor object):
voor *hen* die …, *voor *ze* die …
'for *those* (strong) who …', *'for *those* (weak) who …'

Conversely, the following contexts require a weak pronoun:

(4) Idioms:
dank *je*, *dank *jou*
'thank *you* (weak)', *'thank *you* (strong)'

(5) Generic you:
je weet maar nooit! **jij* weet maar nooit!
'*you* (weak) never can tell!' *'*you* (strong) never can tell'

Generally, strong pronouns are preferred in written language, while weak pronouns are preferred in spoken language, as they are considered more informal. There is a tendency to write strong pronouns which would be weak pronouns in spoken language, and conversely, to pronounce strong pronouns when weak pronouns are read. Strong pronouns sound unnatural when repeated in the same sentence or context. Use of repeated strong pronouns is associated with non-native speakers, since they may be unaware of this unwritten rule. While academic grammars such as Haeseryn et al. (1997) and Donaldson (2008) discuss strong and weak pronouns, the subtleties of their usage are not discussed in most textbooks used by second language learners.

2.2 Linguistic Explanations for the Distribution of Strong/Weak Pronouns

Before going into the linguistic research on the strong/weak pronoun distinction in Dutch, it is helpful to look at research on the production of referring expressions in general. Arnold and Zerkle (2019) investigate why speakers might produce pronouns rather than descriptive noun phrases. Pronouns are strictly less informative than noun phrases, so what other reasons explain their use? The two main explanations they consider are pragmatic and rational factors. The pragmatic model argues that the choice to produce a pronoun can be explained by the speaker's cognitive status of the referent.

Concretely, there is an accessibility hierarchy spanning more or less salient referents. Pronouns are preferred for more salient, recent and frequent referents. The rational model argues that speakers optimize the balance between informativeness and efficiency, with shorter expressions being preferred if they do not cause confusion. Arnold and Zerkle (2019) conclude that the accessibility model only explains part of the variation observed, while efficiency cannot be the primary explanation either. Note also that these theories presuppose that the choice can be explained by rules or efficiency, which is not a given.

Research on strong and weak pronouns also considers the salience hierarchy. Kaiser (2011) summarizes the range of options for referring expressions as follows:

> null > reduced pronoun > full pronoun > demonstrative > noun phrases … etc.
> most salient referent — less salient referent

Kaiser (2011) looks at Dutch specifically. The Dutch language does not have null pronouns except in very limited cases such as imperatives, but it does have reduced (specifically, weak) pronouns, and this distinction is made in both speaking and writing. Weak pronouns are therefore the most salient option available.

In addition to the salience explanation, Kaiser (2011) considers the explanation that strong pronouns are used to express contrast, i.e., the situation where there are multiple competing discourse referents, or where there is a switch to a new topic. She presents data from sentence completion as well as eye tracking experiments. Participants are manipulated using several conditions to test the salience and contrast explanations. The results show that salience does not explain the strong/weak distinction, while it does predict the choice between pronouns and demonstratives. The presence of contrast does result in a marked preference for strong pronouns. However, this does not imply the reverse: that the use of a strong pronoun is likely due to contrast between salient alternatives. This is due to the experimental setup of Kaiser (2011), which has the goal of probing the possible role of referential properties in the strong/weak pronoun distinction; a fortiori, non-referential properties are not considered. Another limitation of the results is that only the strong and weak pronouns *zij/ze* are considered (since first and second person pronouns are not as referentially ambiguous), and only where they occur in subject position (to avoid parallelism effects).

In her general discussion, Kaiser (2011) concludes that the results fit into a form-specific multiple-constraints approach (i.e., there is not a single constraint which can explain the distinction), since uses of strong pronouns that do not express contrast are readily attested. She proposes a Gricean approach in which the use of a strong pronoun where a weak pronoun is also licensed provides an implicature that the strong pronoun was preferred for a reason, such as contrast. The implicature is then further defined to be context-dependent and possibly underspecified (there may not be a

reason). While this account can accommodate any new observation, it does not seem to make any specific, testable predictions.

We can conclude that the linguistic theories underdetermine the data. While speakers are influenced by pragmatics and efficiency, these are not sufficient explanations. We contend that what is missing from these experimental results is a consideration for naturalistic data. The reported experiments create artificial conditions with the goal of testing preconceived hypotheses. This serves to demonstrate that these factors play a role but cannot establish that they are sufficient. We suspect that an overlooked factor is the possibility that strong/weak pronouns also exhibit a stylistic dimension. Especially in the cases where the choice between a strong and weak pronoun is not required or preferred for grammatical reasons, the aforementioned associations of informality and differences in tone may play a role in the selection of a strong rather than a weak pronoun.

3. Study 1: Corpus Frequencies and Correlations

We consider the frequencies of pronoun forms and their correlation with literary ratings. We first look at the frequency of pronouns in general, and then focus on the proportion of strong pronouns in particular.

3.1 Materials and Methods

The corpus consists of 401 contemporary Dutch-language novels by 217 different authors; both originally Dutch and translated novels are included in similar proportions. The novels are best-selling and include different genres, such as thrillers, romantic novels, and literary fiction. In a large survey, readers from the general public rated the literariness of the 401 novels on a seven-point Likert scale (not at all literary to very literary). Survey participants first indicated which novels on the list they had read, and then rated those novels based on the title and author. We use the mean rating per novel as a representative score. While the resulting ratings are ordinal, a sufficient amount of ratings (50–1,000 per book) were collected, and the variance was limited, showing that there was substantial consensus on the literary ratings (for more details, cf. van Cranenburgh et al. 2019).

The texts of the novels were cleaned and automatically parsed with the Dutch *Alpino* parser.[2] The size of the corpus is five million sentences comprising 52 million

2 https://www.let.rug.nl/vannoord/alp/Alpino/

tokens. For our corpus linguistic study, we focus on cases where there may be a choice between weak and strong pronouns, so we only consider pronouns with both versions. We exclude pronouns where the weak version is not common in written language (the neuter pronoun *'t*, female pronoun *d'r*, male pronoun *'m*); we also exclude forms that are exclusively possessive (e.g., *mijn*, *jouw)* and reflexive pronouns (e.g., *mezelf*, *zich)*. This leaves us with the following regular expressions to identify the pronouns of interest (matched at word boundaries, case insensitively):

Strong: `(mij|jij|jou|zij|wij|hen|hun)`
Weak: `(me|je|ze|we)`

After collecting the frequencies of strong and weak pronouns, we determine the correlation with the mean literary rating for each novel.

3.2 Results

We first look at the overall frequency of both pronoun types; for example, more literary novels might focus more on ideas than people. See Figure 1 for the results. There is indeed a negative correlation of pronoun frequency and literariness.

Figure 2 shows the percentage of strong pronouns with respect to both types. In other words, we control for the total number of pronouns, which may differ per novel. Weak pronouns are much more frequent than strong pronouns. On average 82 percent of pronouns with strong and weak forms are weak across the 401 novels. Similarly, in a 700 million word reference corpus (Lassy Large) with edited text from various domains, 76.2 percent of such pronouns are weak.

While we have already found a strong correlation for pronoun frequency, Figure 2 shows that the strong/weak distinction yields a stronger correlation, suggesting that the distinction has a stylistic dimension. A possible explanation would be that weak pronouns are a proxy for informality. However, the result may also be due to more complicated discourse structures in literature which employ strong pronouns for contrast, or a higher frequency of grammatical constructs that require strong pronouns.

Additionally, we see a striking set of nine outliers in Figure 2 with a substantially higher proportion of strong pronouns (>30 percent). The outliers are listed in Table 2. Except for Mitchell, they are novels written by Dutch authors; except for Van Kooten, they are all literary fiction according to the publisher labels and are rated as highly literary by the survey respondents.

Figure 3 shows the distribution of the different pronoun forms in the outliers. The heatmap contrasts the relative frequency (expressed as percentage) of each form with the average relative frequency across the whole corpus of 401 novels with a threshold

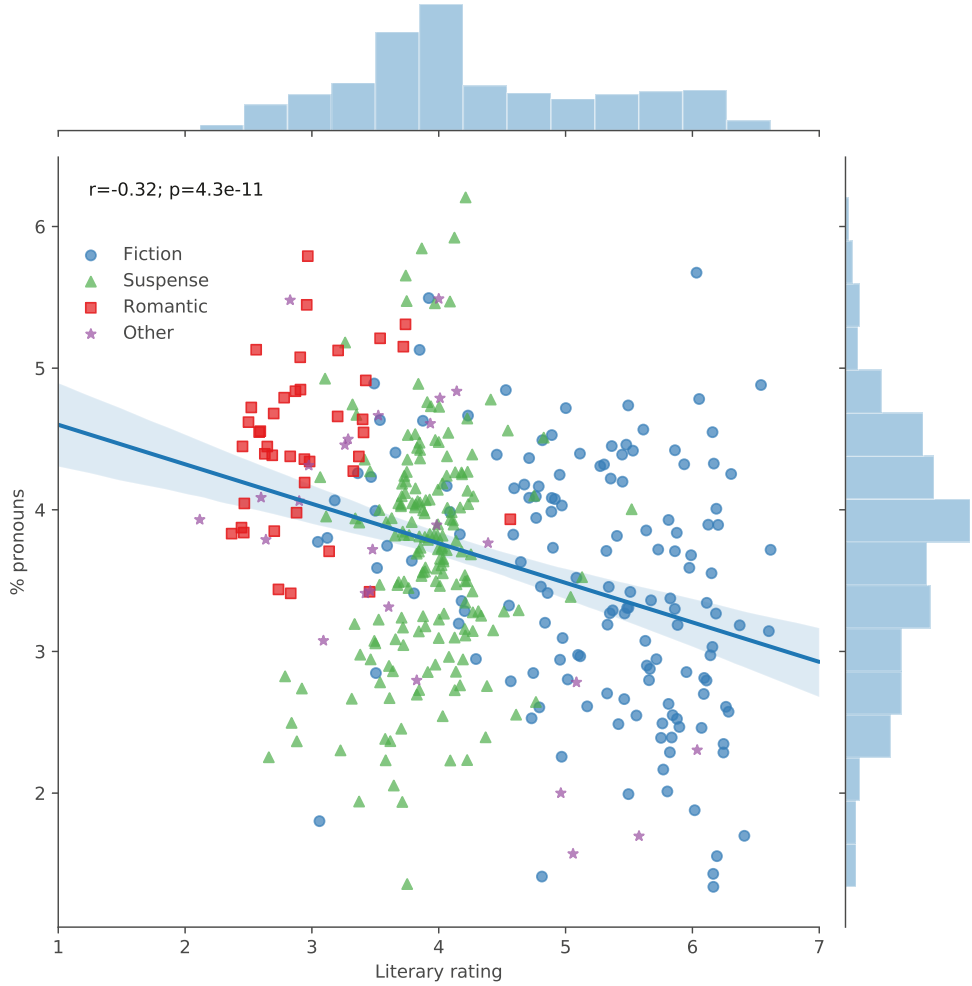

Fig. 1 Percentage of pronouns with respect to all words, correlated against literary ratings (van Cranenburgh, CC BY).

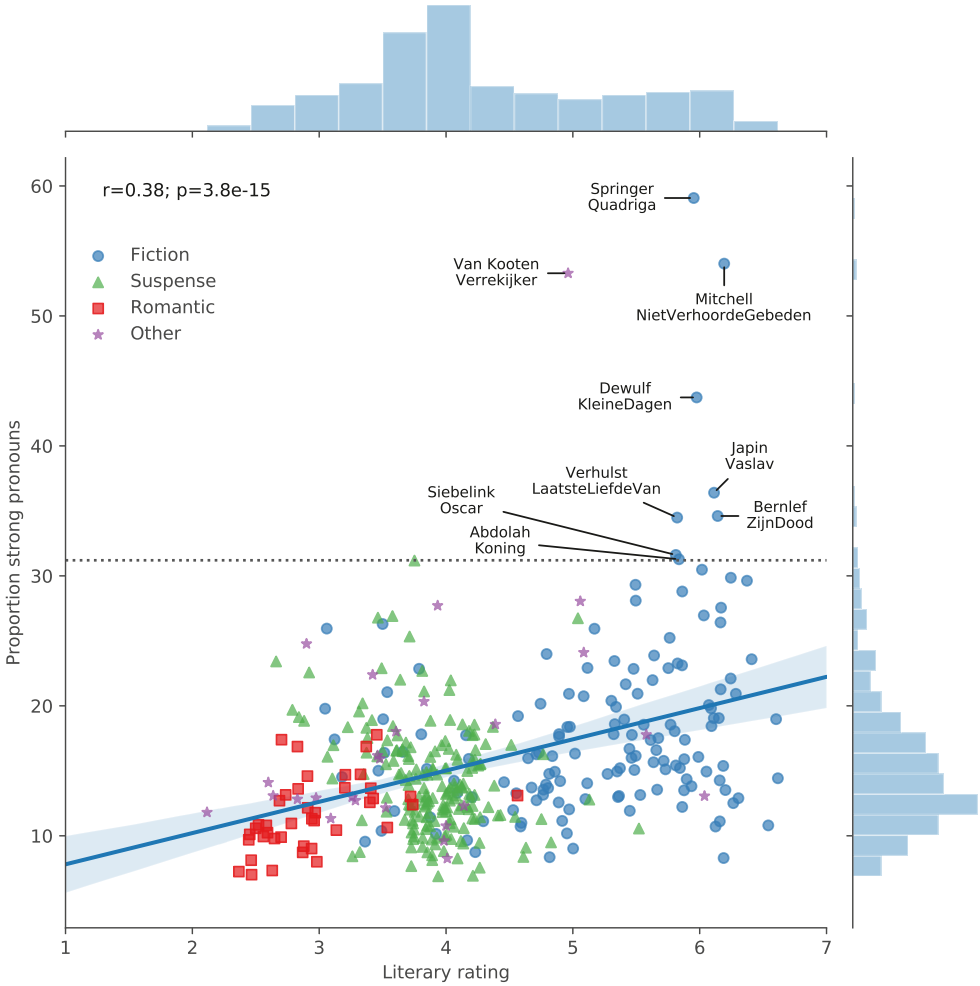

Fig. 2 Percentage of strong pronouns with respect to both pronoun types, correlated against literary ratings (van Cranenburgh, CC BY).

Table 2 Novels that are outliers with respect to the proportion of strong pronouns

	Strong %	Weak %	Both %	Strong prop.
Springer, Quadriga	1.69	1.17	2.85	59.1
Mitchell, The thousand autumns [...]	0.84	0.71	1.55	54.0
Van Kooten, Verrekijker	1.07	0.93	2.00	53.3
Dewulf, Kleine dagen	1.57	2.02	3.59	43.7
Japin, Vaslav	1.22	2.13	3.34	36.4
Bernlef, De een zijn dood	1.03	1.94	2.97	34.6
Verhulst, De laatste liefde van [...]	0.79	1.50	2.29	34.5
Siebelink, Oscar	0.83	1.80	2.63	31.6
Abdolah, Koning	0.75	1.64	2.39	31.3

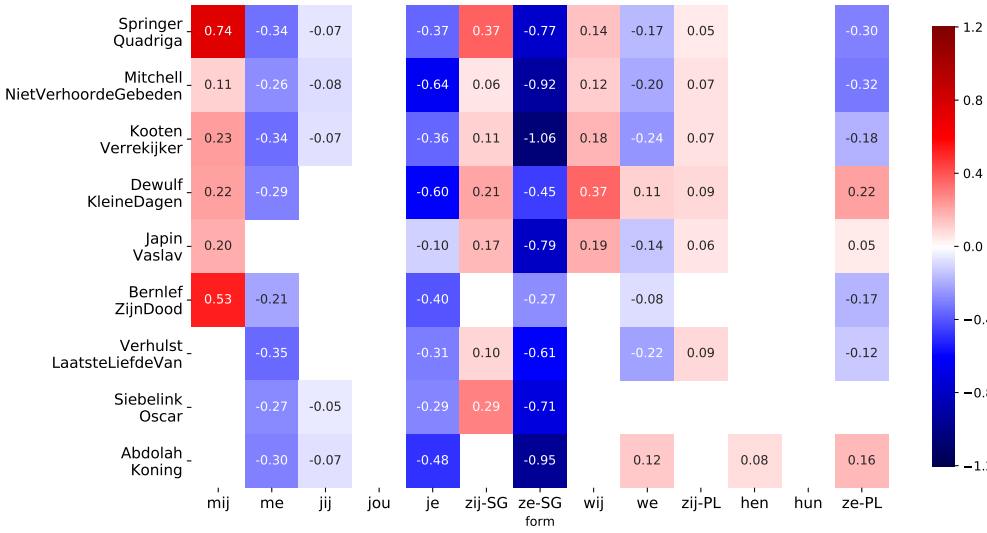

Fig. 3 Heatmap showing the divergence in frequency of the different pronoun forms across the outliers compared to the whole corpus (van Cranenburgh, CC BY).

of 0.05 absolute difference in relative frequency. We used part-of-speech (POS) tags to restrict the counts to occurrences of personal and personal/reflexive pronouns (specifically, VNW(pers,…) and VNW(pr,…); this excludes possessive pronouns and the archaic verb *zij* which were not excluded with the query used for Figure 2). Since the pronoun *zij/ze* can be both feminine/singular and plural, these are listed separately. We find that certain forms do not differ appreciably (*jou*, *hun*) in any of the novels, while this holds for the forms *mij* and *wij* in three and four novels, respectively. A few strong pronouns are shown in dark red, indicating that they are much more frequent than average; many more weak pronouns are shown in dark blue, indicating that they are much less frequent than average in these outlier novels.

Now that a corpus study has revealed these outlier novels, we will attempt to explain their outlier status using the linguistic contexts in which the pronouns are used.

4. Study 2: Coding Contrast and Preference

We now take a closer look at the outliers. While a comprehensive study of the observed variation would need to contrast the outliers with non-outliers, we focus here on investigating why the outliers display such an exceptionally high proportion of strong pronouns. We will consider whether any of the proposed reasons for the use of strong and weak pronouns apply: contexts in which one or the other is grammatically obligatory and contexts which may lead one or the other to be preferred.

4.1 Materials and Methods

We annotated the relevant strong/weak pronouns in the first 100 sentences of each of the nine outlier novels, resulting in 356 annotated pronouns. We used the following coding scheme:

- Pronoun form (strong versus weak)
- In this sentence, is the given pronoun:
  - Free choice (both strong and weak pronoun seem equally acceptable)
  - Preferred (the other pronoun would be dispreferred)
  - Obligatory (the other pronoun would be ungrammatical)
- Is the pronoun used for contrast (yes/no)?
- Pronoun type/POS: personal, possessive, generic, impersonal, verb (the form *zij* is also an archaic form of the verb to be)

The annotations were done by a single annotator (the present author), so no inter-annotator agreement score can be estimated. However, the binary distinction between grammatically obligatory or not is clear cut. The distinction between free choice and preferred is admittedly more subjective and judgments from multiple annotators would improve reliability; moreover, it might be better conceived as a spectrum of acceptability. However, the distinction is important for our research question, since the proposed explanations by Kaiser (2011) concern uses of pronouns where both forms are possible, but one is preferred. For the decision whether contrast was present, no context beyond the sentence was considered. Judging from the examples cited by Kaiser and Trueswell (2004), our notion of contrast is stricter; consider Kaiser and Trueswell (2004, 146):

(6) [context: Gilles and Ange are in the kitchen, and Ange notices Gilles looking outside intently. She tries to look [as well], knowing that outside are a garden, a river, the whole world.]
[…] maar hoe Ange zich ook inspant om van dat alles een glimp te ontwaren, *zij* ziet in de donkere ruit slechts de weerspiegeling van haar eigen keuken […] (Dorrestein, *Het hemelse gerecht*, p. 15)
'[…] but no matter how Ange exerts herself trying to catch a glimpse of all that, *she* sees in the dark pane nothing but her own kitchen […]'

If this pronoun is to be judged contrastive, this comes implicitly from the context and is a matter of interpretation; the pronoun has no emphasis or focus, and a weak pronoun would arguably fit equally well. Compare this example to the overt examples of contrast in the reference grammar *Algemene Nederlandse Spraakkunst* which require no context (Haeseryn et al. 1997, 252):

(7) Hij bedoelt *jou* niet, maar Mark.
'He does not mean *you*, but Mark.'

(8) Ik vind jouw verhaal veel geloofwaardiger dan dat van *hem*.
'I find your story much more credible than *his* one.'

Still, in (8) the contrast can also be on *jouw* (your) instead of *hem (*his). It seems difficult to operationalize the notion of contrast rigorously, and to avoid confirmation bias during annotation (for both presence and absence of contrast).

4.2 Results

Table 3 lists the distribution of pronoun types. For our research question, we focus on the personal pronouns, since the other types do not allow for both forms.[3] The personal pronouns also form the majority; we can therefore rule out that the other types are responsible for the outliers.

Table 3 Distribution of types

Personal	303
Possessive	26
Generic	20
Impersonal	6
Verb	1
Total	**356**

We will continue the analysis with only the 303 personal pronoun tokens. Figure 4 shows the breakdown of the other coded variables. We can conclude that the use of emphasis/contrast is rare. It can therefore be ruled out as an explanation for the outlying novels.

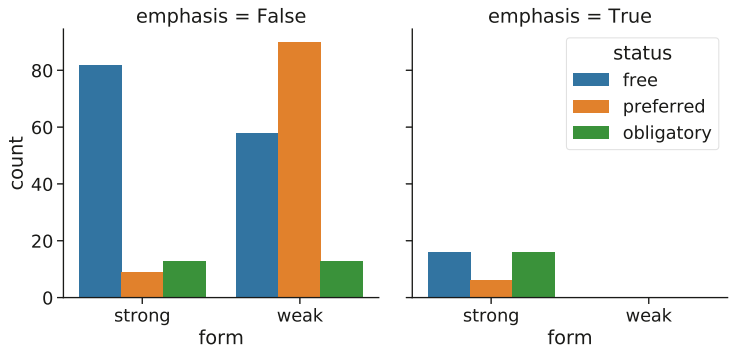

Fig. 4 Breakdown of manually analyzed pronouns (N = 303) (van Cranenburgh, CC BY).

When we focus on the pronouns without emphasis, which form the majority, we see that weak pronouns are often preferred, while strong pronouns are rarely preferred. Overall, a large proportion of both strong and weak pronouns are free choice, without preference for either form. This means that the grammatical explanations of salience and contrast cannot explain our observed outliers. Moreover, these results support the hypothesis that a large part of the strong/weak distinction is a stylistic matter. The following are typical examples of each category:

3 The possessive *je* has the strong form *jouw*, but we choose to focus on personal pronouns.

(9) Weak pronouns:
  a. Free: *We* speuren erfgenamen op
    '*We* track down heirs'
  b. Preferred: Dat weet *je* toch?
    '*You* know that right?'
  c. Obligatory: Mooie gouvernante is *me* dat.
    'Nice governess that is.'

(10) Strong pronouns:
  a. Free: Hoort u *mij?*
    'Do you hear *me*?'
  b. Preferred: Maar dan kennen ze *mij* niet.
    'But then they haven't met *me*.'
  c. Obligatory: Je ziet dat het niet van *mij* is!
    'You can tell it's not *mine*!'

We also encountered some particularly interesting examples. The following are arguably unnatural usage of a strong reflexive pronoun, which may have been chosen for deliberate stylistic effect:

(11) a. Ik keek om *mij* heen
    'I looked around *me*'
  b. aangezien […] heb ik altijd mijn eigen Duralexglas bij *mij*
    'since […] I always have my own Duralex glass with *me*'

In the following sentence, we have a clear example of a strong pronoun expressing contrast, as the strong pronoun picks out a different referent than the weak pronoun which occurs in the same sentence.

(12) Ik heb nooit kunnen vaststellen dat *ze* mij in de gaten hielden, al deden *ze* dat natuurlijk wel, en *zij* in de eerste plaats.
  'I have never been able to confirm that *they* were watching me, although of course *they* did , and *she* most of all.'

However, it should be noted that this was the only clear example of contrast in the 303 pronouns we coded. This finding strongly contrasts with the preliminary corpus study of Kaiser and Trueswell (2004), who report that most uses of the strong pronoun *zij* are prompted by contrast.

5. Study 3: Coreference Analysis

A limitation of the approach in the previous section is the amount of subjectivity involved in the annotation. We now consider a more clearly defined task: coreference annotation. Given a pronoun, it is a well-defined question what other expressions in the text refer to the same entity. Coupled with the parse trees of sentences, we can directly test some of the proposed explanations for the distribution of strong and weak pronouns.

5.1 Materials and Methods

Fragments of a selection of 33 novels were manually annotated for coreference using the annotation scheme described in van Cranenburgh (2019). The set includes both the outliers as well as different kinds of novels without a high proportion of strong pronouns. The length of the fragments ranges from 1,000 to 20,000 tokens, rounded to the nearest sentence boundary. In total the subcorpus annotated for coreference contains 172,544 tokens. From the coreference annotations we extract the following predictors for each strong/weak pronoun:

1. pronoun function (core or non-core): non-core (i.e., not subject or object) arguments tend to be strong pronouns.
2. antecedent function (subject or other): the grammatical function of the antecedent (i.e., the closest preceding mention); to avoid sparsity, we only use an indicator for whether the antecedent is a subject or not. Subjects are more prominent and therefore more likely to be referred to by a strong pronoun.
3. distance: the distance to the antecedent in number of sentences. The distance is log transformed since it has a skewed distribution (most antecedents are close). A recently mentioned referent tends to be referred to by a weak pronoun.
4. chain density: the number of mentions in the same coreference chain as the pronoun in a window of 10 preceding sentences. A frequently used referent tends to be referred to by a weak pronoun.

These independent variables are compared to the dependent variable, whether a pronoun is strong or weak. We have also considered the number of competing mentions between the antecedent and the pronoun, but this was strongly correlated ($r = 0.83$) with distance and is therefore left out.

5.2 Results

A logistic regression with N = 3,549 strong/weak pronouns (Table 4) shows that except for distance, these variables are significant predictors.

Table 4 Logistic regression predicting strong pronouns from several proposed predictors

Dependent variable		strong vs weak pronoun	
No. observations		3,549	
Pseudo R-squ.		0.04412	
LLR p-value		1.650e-35	
	coef	std err	p
(intercept)	0.4882	0.157	0.002
pronfunc = core	−1.5708	0.128	0.000
antfunc = subj	−0.2851	0.082	0.001
log(distance)	0.0830	0.045	0.063
chain density	−0.0299	0.009	0.001

The continuous predictors are visualized in Figure 5. All signs of the coefficients match the hypothesized explanations: a negative coefficient indicates the variable makes a strong pronoun less likely, and vice versa. The logistic regression as a whole has a significant log-likelihood ratio as well, but the pseudo-R^2 is low. Since pseudo-R^2 is hard to interpret, we also calculate the area under the ROC-curve (a.k.a. concordance index); we compute this without cross-validation (within sample) and find C = 0.611. According to Hosmer and Lemeshow (2000, 162), C = 0.5 means no discrimination,

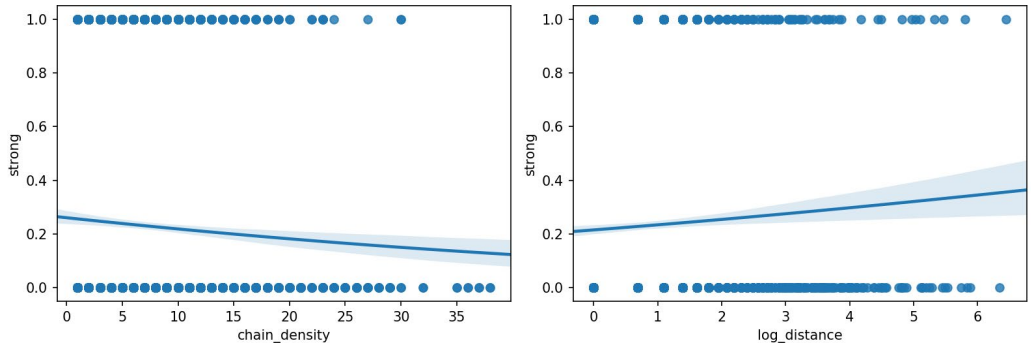

Fig. 5 Logistic regression plot of probability of a strong pronoun against chain density and distance to antecedent (van Cranenburgh, CC BY)

while $0.7 \leq C < 0.8$ means acceptable discrimination. We conclude that the model fit using these predictors is weak.

6. Discussion and Conclusion

We return to our research questions: we discovered a striking association between the use of strong/weak pronouns and literariness. There is a negative correlation with the number of pronouns and literariness, an expected result since both pronouns and less literary novels are associated with informal and spoken language. On the other hand, there is a positive correlation with the proportion of strong pronouns and literariness. A set of nine literary novels have a much larger than average proportion of strong pronouns. A manual analysis shows that these novels are not outliers due to grammatically obligatory strong pronouns, and a preference for strong pronouns is rare. Ergo, their authors freely chose to use a large number of strong pronouns, without being prompted by any of the proposed discourse-related factors. This was again confirmed by the statistical analysis of coreference annotations. The linguistic explanations for the use of strong and weak pronouns are shown to be significant variables but the amount of variance explained is limited; moreover, they cannot explain the outliers. What remains as a likely explanation is a stylistic dimension, given that choice is involved (whether the choice is deliberate is a second question). We submit that style is an important aspect of the use of strong and weak pronouns and referring expressions in general. Future research should investigate the stylistic effects of strong and weak pronouns in more detail by collecting fine-grained judgments from multiple annotators. Specifically, we should establish more precisely the degree of freedom in particular contexts, and the perceived unnaturalness of a variety of observed examples from both outliers and ordinary novels. The latter may turn out to be a clear instance of defamiliarization in literary language.

Code and data repository

https://github.com/andreasvc/strongweaklit

Acknowledgments

The author wishes to acknowledge helpful comments on this paper from Remco Knooihuizen, Jack Hoeksema, Kim Jautze, and an anonymous reviewer.

ORCID®

Andreas van Cranenburgh https://orcid.org/0000-0002-4545-1548

References

Arnold, Jennifer E., and Sandra A. Zerkle. 2019. "Why Do People Produce Pronouns? Pragmatic Selection vs. Rational Models." *Language, Cognition and Neuroscience,* 34 (9): 1152–75. https://doi.org/10.1080/23273798.2019.1636103.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Bresnan, Joan. 1998. "Markedness and Morphosyntactic Variation in Pronominal Systems." In *Workshop Is Syntax Different.* http://web.stanford.edu/~bresnan/wow98-8.ps.

Donaldson, Bruce. 2008. *Dutch: A Comprehensive Grammar*. London: Routledge.

Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jacobus De Rooij, and Maarten Cornelis van den Toorn, eds. 1997. *Algemene Nederlandse Spraakkunst* [General Dutch Grammar]. Groningen: Martinus Nijhoff. https://e-ans.ivdnt.org/.

Hosmer, David W., and Stanley Lemeshow. 2000. *Applied Logistic Regression*. New York: Wiley.

Kaiser, Elsi. 2011. "Salience and Contrast Effects in Reference Resolution: The Interpretation of Dutch Pronouns and Demonstratives." *Language and Cognitive Processes* 26 (10): 1587–1624. https://doi.org/10.1080/01690965.2010.522915.

Kaiser, Elsi, and John Trueswell. 2004. "The Referential Properties of Dutch Pronouns and Demonstratives: Is Salience Enough." *Proceedings of Sinn und Bedeutung* 8 (August), 137–50: Konstanz: University of Konstanz. https://doi.org/10.18148/sub/2004.v8i0.754.

Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey." *Poetics* 79:101439. https://doi.org/10.1016/j.poetic.2020.101439.

van Cranenburgh, Andreas. 2019. "A Dutch Coreference Resolution System with an Evaluation on Literary Fiction." *Computational Linguistics in the Netherlands Journal* 9: 27–54. https://clinjournal.org/clinj/article/view/91.

van Cranenburgh, Andreas, and Rens Bod. 2017. "A Data-Oriented Model of Literary Language." In *Proceedings of EACL*, 1228–38, Association for Computational Linguistics. http://aclweb.org/anthology/E17-1115.

van Cranenburgh, Andreas, Karina van Dalen-Oskam, and Joris van Zundert. 2019. "Vector Space Explorations of Literary Language." *Language Resources and Evaluation* 53: 625–50. https://doi.org/10.1007/s10579-018-09442-4.