The Strong-Weak Pronoun Distinction as a Marker of Literariness

Andreas van Cranenburgh

28 February 2019, Würzburg
Outline

- General problem: What is literature?
- This talk: Strong/weak pronouns
- Results
  - Correlations
  - Outliers
  - Analysis
Background: What is literature?

What makes a literary novel *literary*?

- Cultural capital of critics and publishers
- Subjective aesthetic value-judgments
- “Objective” textual features, writing style: *literariness*
The Riddle of Literary Quality

- 401 contemporary Dutch-language novels
- Large reader survey of general public
- Literary ratings on 7-point Likert scale
- Stylometry and machine learning with texts of novels

http://literaryquality.huygens.knaw.nl
Previous results: 76.0 % $R^2$

- Literariness is highly predictable from text
- Word frequencies, cliches, syntactic complexity yield good predictions,
- but many many features, hard to interpret.

This talk is not about getting better predictions, but about understanding one specific stylistic aspect

van Cranenburgh et al (2019). Vector space explorations of literary language
Some Dutch pronouns have strong and weak (reduced) forms

- Same meaning, but strong/weak is sometimes obligatory or preferred
- Other times, free choice (style!)

<table>
<thead>
<tr>
<th></th>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj, obj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st sg</td>
<td>ik, mij</td>
<td>me</td>
</tr>
<tr>
<td>2nd sg</td>
<td>jij, jou</td>
<td>je</td>
</tr>
<tr>
<td>3rd sg fem</td>
<td>zij, haar</td>
<td>ze</td>
</tr>
<tr>
<td>3rd sg masc</td>
<td>hij, hem</td>
<td>-</td>
</tr>
<tr>
<td>3rd sg neut</td>
<td>het</td>
<td>-</td>
</tr>
<tr>
<td>1st pl</td>
<td>wij, ons</td>
<td>we</td>
</tr>
<tr>
<td>2nd pl</td>
<td>jullie</td>
<td>-</td>
</tr>
<tr>
<td>3rd pl</td>
<td>zij, hen/hun</td>
<td>ze</td>
</tr>
</tbody>
</table>

Red/blue: pronouns with both forms
Not shown: weak pronouns avoided in written language or only used as possessive/reflexive.

Cf. Haeseryn et al. (1997). Algemene Nederlandse Spraakkunst
Previous work: why strong vs weak?

Kaiser (2011) and works cited there:

**Salience:**

null > reduced pronoun > full pronoun > demonstrative > full NP ... etc.

most salient referent

less salient referent

**Contrast:**

The referent is in a contrast relation to other entities in the discourse

This talk

Missing from previous work:

**Stylistic dimension:**
Weak vs strong pronouns are related to informality and the tone of a text
This talk

Missing from previous work:

Stylistic dimension:
Weak vs strong pronouns are related to informality and the tone of a text

Research Questions:

1. Is there an association between strong/weak pronouns and literariness?
2. How common is it for these variants to be
   a. Free stylistic choice
   b. Preferred
   c. Obligatory
3. Why could there be such an association?
Method

Calculate correlation between:

1. Mean literary ratings of novels
2. Two independent measurements:
   a. Baseline: frequency of both pronoun forms
   b. Proportion of strong pronouns vs both forms
Correlation: pronouns vs literary rating (N=401)

- Count both forms
- Divide by total words (rel. freq)
- Result: Less pronouns, more literary
- Probably proxy for amount of dialogue vs narrative description
- Count strong pronouns, divide by count of both forms
- Independent of total number of pronouns
- On average, 85% of pronouns are weak
- More strong pronouns, more literary
- Several strong outliers!
Table: Divergence of relative frequencies (wrt corpus mean)

- All Dutch authors (except Mitchell), highly literary (> 5)
- Less je, ze-SG,FEM. More mij, wij, zij-SG,FEM
Possible explanations

Why are strong pronouns more common in literary texts?

**Stylistic choice** (deliberate or not):
- Non-literary texts have more informal, idiomatic language
- Literary authors are less afraid of sounding “unnatural”

**Discourse structure** more complicated:
- Larger number of characters
- Multiple perspective, storylines give rise to higher frequency of less salient referents and use of contrast.
Manual Analysis

- In first 100 sentences of the outliers
- Annotate each pronoun:
  - Strong vs weak
  - Free choice, preferred, or obligatory
  - Used for emphasis/contrast?
  - Type: personal, possessive, generic, non-personal, verb

Limitation: annotation was done without looking at discourse context.
Manual Analysis

- In first 100 sentences of the outliers
- Annotate each pronoun:
  - Strong vs weak
  - Free choice, preferred, or obligatory
  - Used for emphasis/contrast?
  - Type: personal, possessive, generic, non-personal, verb

Limitation: annotation was done without looking at discourse context.

Distribution of types:
- Personal 303
- Possessive 26
- Generic 20
- Non-personal 6
- Verb 1
- Total 356

We’ll only consider personal pronouns, which allow both forms.
Emphasis is rare.

Weak often preferred, but large part is free choice.
Typical examples

Weak:

(1) 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.*

Strong:

(2) 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!*
Interesting examples

Arguably unnatural usage of strong pronoun:

(3) a. Ik keek om mijn heen
    *I looked around me*

    b. aangezien [...] heb ik altijd mijn eigen Duralexglas bij mijn
    *since [...] I always have my own Duralex glass with me*

Weak vs strong pronouns pick different referents:

(4) 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.*
Conclusion

Answers to Research Questions:

1. Is there an association between strong/weak pronouns and literariness?
   - Negative correlation between pronouns and literariness
   - Positive correlation between strong pronouns and literariness
   - Striking outliers: Dutch literary authors who use lots of strong pronouns
Conclusion

Answers to Research Questions:

1. Is there an association between strong/weak pronouns and literariness?
   - Negative correlation between pronouns and literariness
   - Positive correlation between strong pronouns and literariness
   - Striking outliers: Dutch literary authors who use lots of strong pronouns

2. How common is it for these variants to be free choice, preferred, or obligatory?
   - Majority is free, stylistic choice.
   - Even in the outliers, weak pronouns are often preferred
   - Obligatoriness and emphasis are rare.
Answers to Research Questions:

1. Is there an association between strong/weak pronouns and literariness?
   - Negative correlation between pronouns and literariness
   - Positive correlation between strong pronouns and literariness
   - Striking outliers: Dutch literary authors who use lots of strong pronouns

2. How common is it for these variants to be free choice, preferred, or obligatory?
   - Majority is free, stylistic choice.
   - Even in the outliers, weak pronouns are often preferred
   - Obligatoriness and emphasis are rare.

3. Why could there be such an association?
   - Seems to be predominantly a stylistic choice
   - Should look into discourse context
EXTRA SLIDES
Future Work

▶ Analyze antecedents: distance, grammatical function, competing referents
▶ Compare with non-literary corpora (news etc.)
▶ See whether weak/strong distinction is useful as a feature for automatic anaphora resolution
▶ Analyze dialogue for each character separately from narrative text
Properties of strong and weak pronouns

**Strong:** (emphatic)
- Default, unmarked
- Often stressed: emphasis/contrast
- Less salient referents
- Obligatory for:
  - Comparisons
    (ik ben rijker dan jij, *dan je)
  - Conjunctions of two pronouns
    (hij en zij)
  - Oblique arguments, eg. rel. clause (voor hen die . . .)
- Preferred:
  - Sounds unnatural when repeated
  - Preferred in writing, even when weak form used when same sentence is spoken (applies to ie, ’m, d’r, ’t, but also to a lesser extent for me/je/we/ze)

**Weak:** (unemphatic)
- Reduced, marked
- Always unstressed: no emphasis
- Salient referents
- Obligatory for:
  - Idioms
    (e.g., dank je, *dank jij)
  - Generic you (je weet maar nooit!)
  - 3rd pers. pl. non-personal
    (ze, *zij, *hun, *hun);
- Preferred:
  - Can be repeated; or sentence has one strong pronoun followed by several weak instances
  - Less personal, more informal

Cf. Haeseryn et al. (1997). Algemene Nederlandse Spraakkunst; Bresnan (1998), Markedness and morphosyntactic variation in pronominal systems
Distribution of grammatical functions

Percentages add up to 100 for each row