Beyond the textual company of words: What corpus settings tell us about lexical collocability

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

[1] Leuven University College
[2] KU Leuven, Quantitative Lexicology and Variational Linguistics

IVACS 2014 – Newcastle, 19-21/6/2014

Contents

1. Problem statement

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Lexical collocations

- Long-standing tradition in corpus linguistic research, dating back to 50ies (amongst others, Firth 1957; Granger 1998; Hoey 2005; Sinclair 1991; Stubbs 1995, 2001; Wulff 2008, 2013; see Gries 2013 for critical methodological account)

- Use in its own right to identify lexical preference patterns, in various linguistic disciplines

- Use as explanatory variable / determinant to constrain other constructions

1. Problem statement

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Corpus (1/2)

- Representative sample of **language use** of a given linguistic community in a/given **setting(s)**

- Corpus-based approaches: focus on linguistic patterns and structures in language use

- Settings of language use:
  - Rarely explicitly addressed in mainstream (corpus) linguistics
  - Object of peripheral linguistic disciplines (sociolinguistics, dialectology, stylistics, etc.)

1. Problem statement

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Corpus (2/2)

- Settings of language use: reflection of
  - Variety of usage settings
  - Heterogeneity linguistic community

  (Heylen et al. 2008)

  ] socio-cultural diversity

- Research lexical collocation: impact of language settings hardly explicitly addressed
  (exception: Stefanowitsch & Gries 2008)

2. Goal

Demonstrate that lexical collocations are subject to constraints from usage settings

1. as measures in their own right to identify lexical preference patterns

2. as explanatory variables

Procedure: case study

3. Case study

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Adjectival inflection in Dutch definite NPs with singular neuter N_{head}

- Two alternating morphosyntactic realizations:
  - [**inflected**] -e *het vriendelijk-e kind* ('the friendly-INFL child')
  - [**uninflected**] -Ø *het vriendelijk-Ø kind* ('the friendly-ZERO child')

- **Alternation** governed by intricate network of explanatory variables (Haeseryn et al. 1997; Tummers 2005)
  - Structural: lexical collocation strength AN, Det_{POS}, N_{dim}, N_{inf}, …
  - Usage settings: national variety, register
  - Discourse processing: prosodic pattern AN

- Present talk: focus on
  - Lexical collocation strength AN
  - Register
  - National variety
  - Speaker

3. Case study

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Corpus

- Corpus of spoken Dutch (*Corpus Gesproken Nederlands*; Oostdijk 2000)
  - 10M reference corpus of spoken Dutch
  - National variety: Belgian Dutch vs. Netherlandic Dutch
  - Register: different degrees of speaker control on situation

- **Corpus distribution** adjectival alternatives

	n	%
Inflected	3,810	0.7675
Uninflected	1,154	0.2325
Total	**4,964**	**1.000**

3. Methodology

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Operationalization of variables (1/5)

- `lex.col`:
  - Lexical collocation strength between A and N (in NP)
  - Pointwise mutual information index (Church & Hanks 1990)
    - Computed based on lemmas in *Leuven News Corpus* (1.3 billion words; Ruette 2012) and *Twente News Corpus* (560 million words; Ordelman et al. 2007) for AN pairs
    - Transposed to dataset

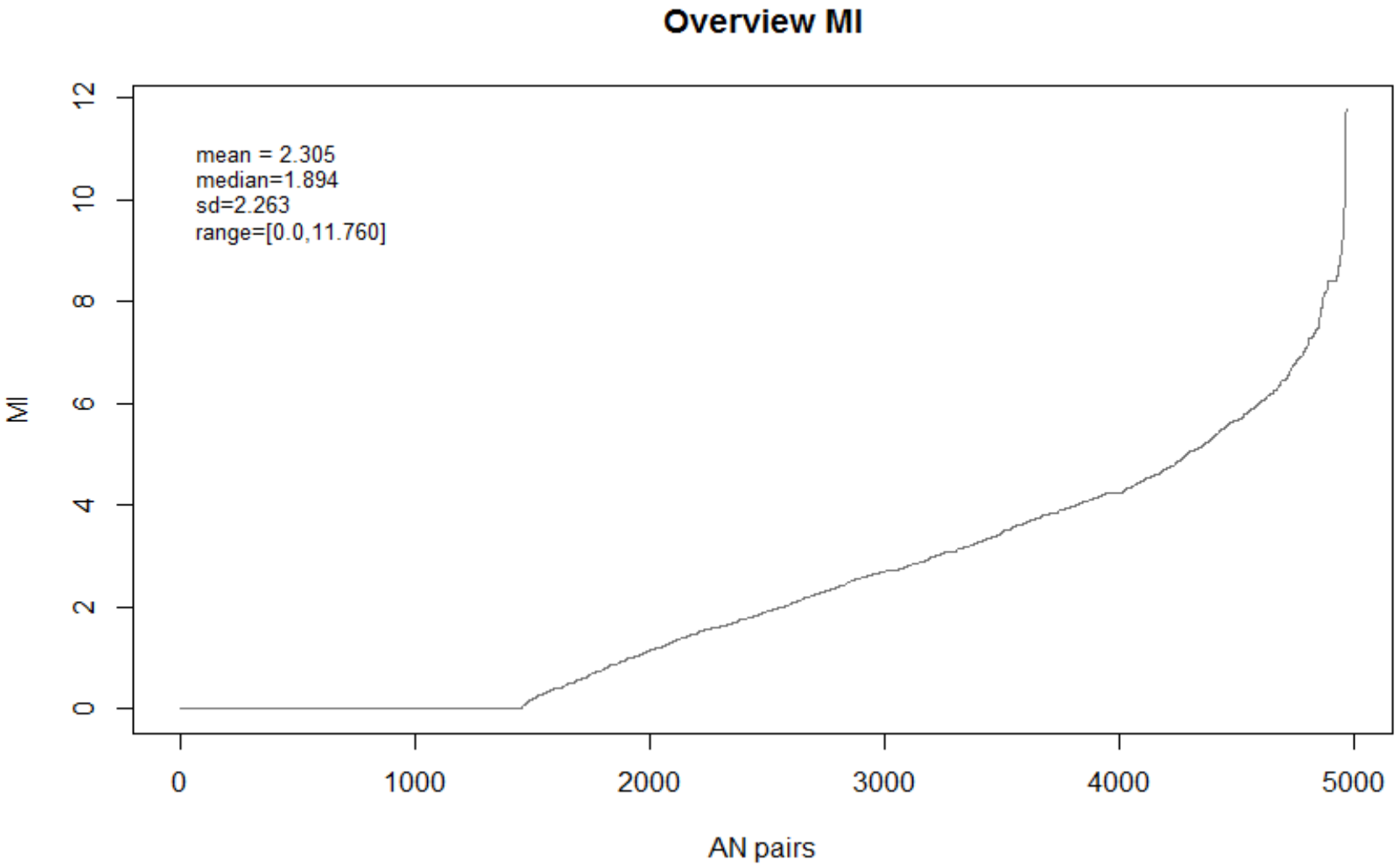
- `nat.var`: Netherlandic vs. Belgian Dutch

- `register`:
  - `high.form` > `mod.form` > `mod.inf` > `high.inf`
  - Based on 3 binary stylistic dimensions in CGN
    - preparation: prepared vs. non-prepared
    - audience: public vs. private
    - interaction: monologue vs. dia- or multilogue

3. Methodology

Operationalization of variables (2/5)

- `lex.col`:



-
-

3. Methodology

Operationalization of variables (3/5)

- speaker:
  - Assumption of independence of observations: often violated in corpora
  - Observations are ① grouped under speakers, ② who will (probably) be different in replication studies

  - **Problems**
    - ① Grouping
      - Speakers' idiosyncratic tendencies
      - Size of speaker's contribution
    - ② Generalizability

3. Methodology

Operationalization of variables (4/5)

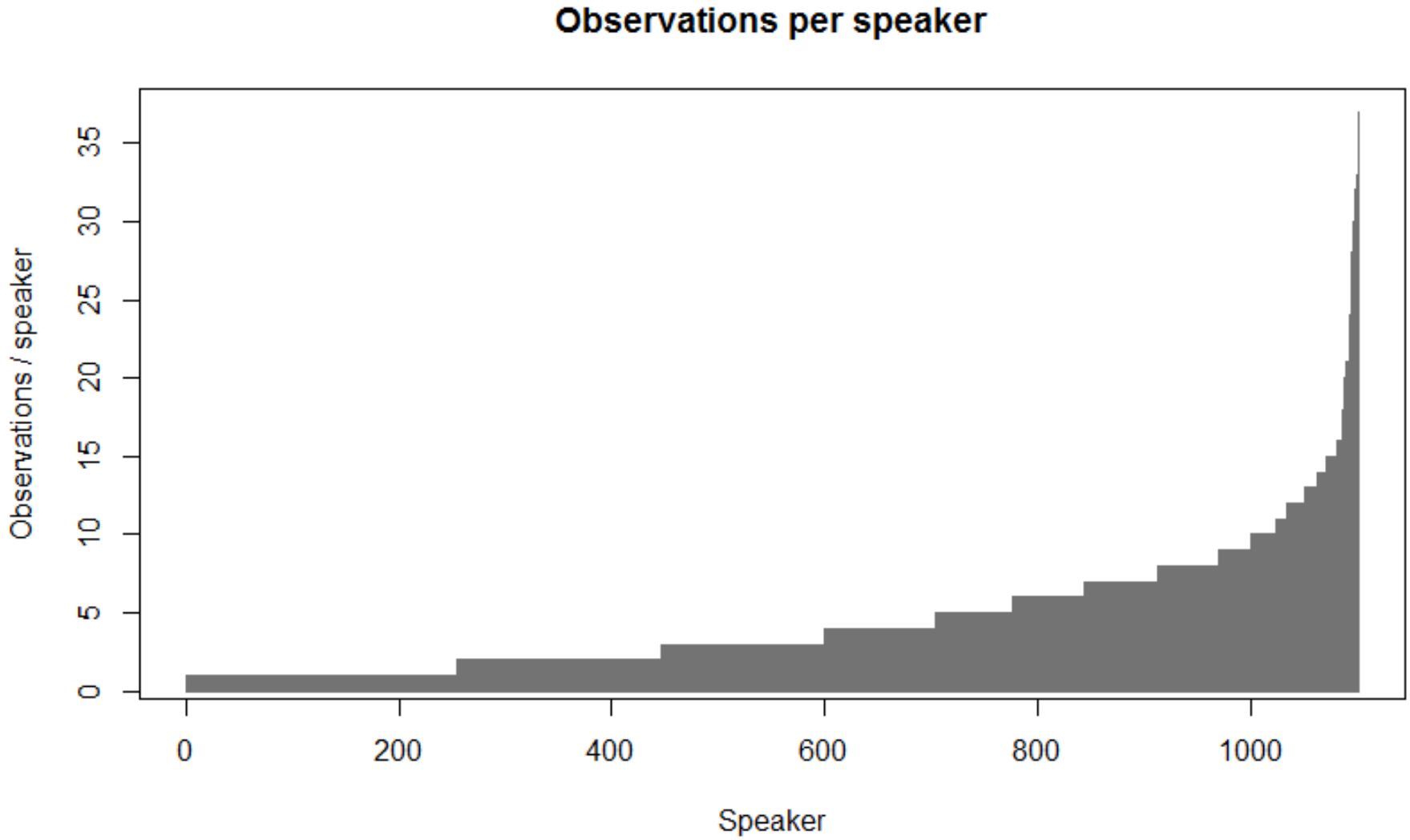
- `speaker`: overview statistics

	Speakers	Observations
Single contributor	253 (0.23)	253 (0.05)
Multiple contributor	848 (0.77)	4,711 (0.95)
Total	1,101 (1.00)	4,964 (1.00)

3. Methodology

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Operationalization of variables (5/5)



4. Methodology

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Modeling: mixed-effects models (1/2)

- **Fixed effect terms**: exhaust all levels of parameter; identical values in replication study
  - `lex.col`
  - `nat.var`
  - `register`

- **Random effect term**: sampled from larger population; different values in replication study
  - `speaker`

(Baayen 2008; Bates & Pinheiro 2000; Gelman & Hill 2007)

4. Methodology

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Modeling: mixed-effects models (2/2)

- **Modeling lexical collocation strength**:

```
lex.col ~ register * nat.var                    [fixed]
          + (1 | speaker)                       [random]
```

- **Modeling adjectival inflection:**

$$\ln(^{A.uninfl}/_{A.nfl}) \sim lex.col * register * nat.var + (1 + lex.col \mid speaker)$$

[fixed] [random]

- **Analyses**: R
  - `lme4` library (Bates 2005; Bates et al. 2013)
  - `arm` library (Gelman & Hill 2007)
  - `effects` library (Fox 2008)
  - `car` library (Fox & Weisberg 2011)

5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

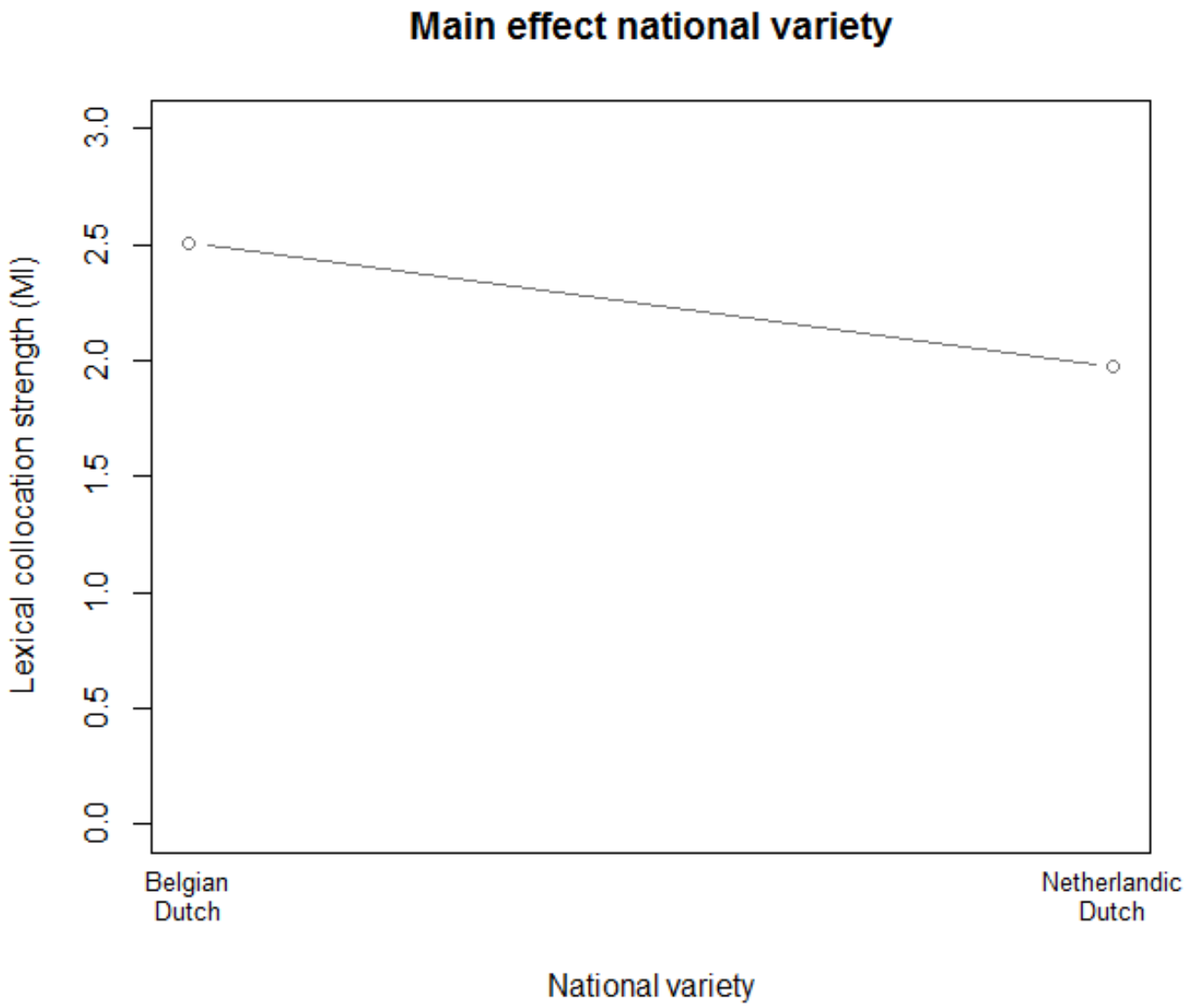
Collocation strength AN pair

- Model summary: sequential anova (Fox 2008)

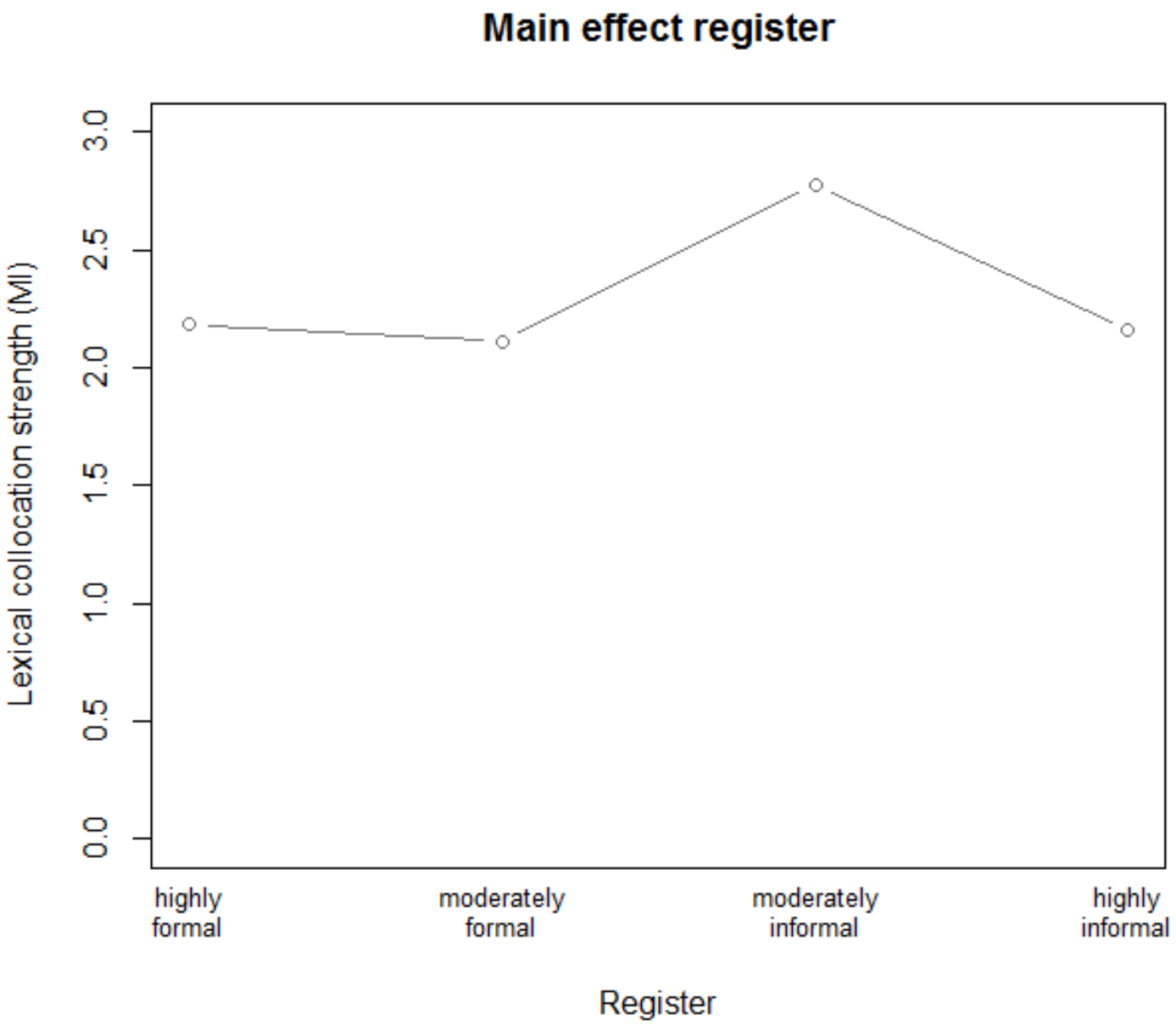
```
Analysis of Deviance Table (Type II Wald chisquare tests)

Response: lex.col
                  Chisq Df Pr(>Chisq)
nat.var          28.217  1  1.085e-07 ***
register         37.080  3  4.426e-08 ***
nat.var:register 12.484  3   0.005895 **
```

- Overview fixed effects and random effect (speaker)

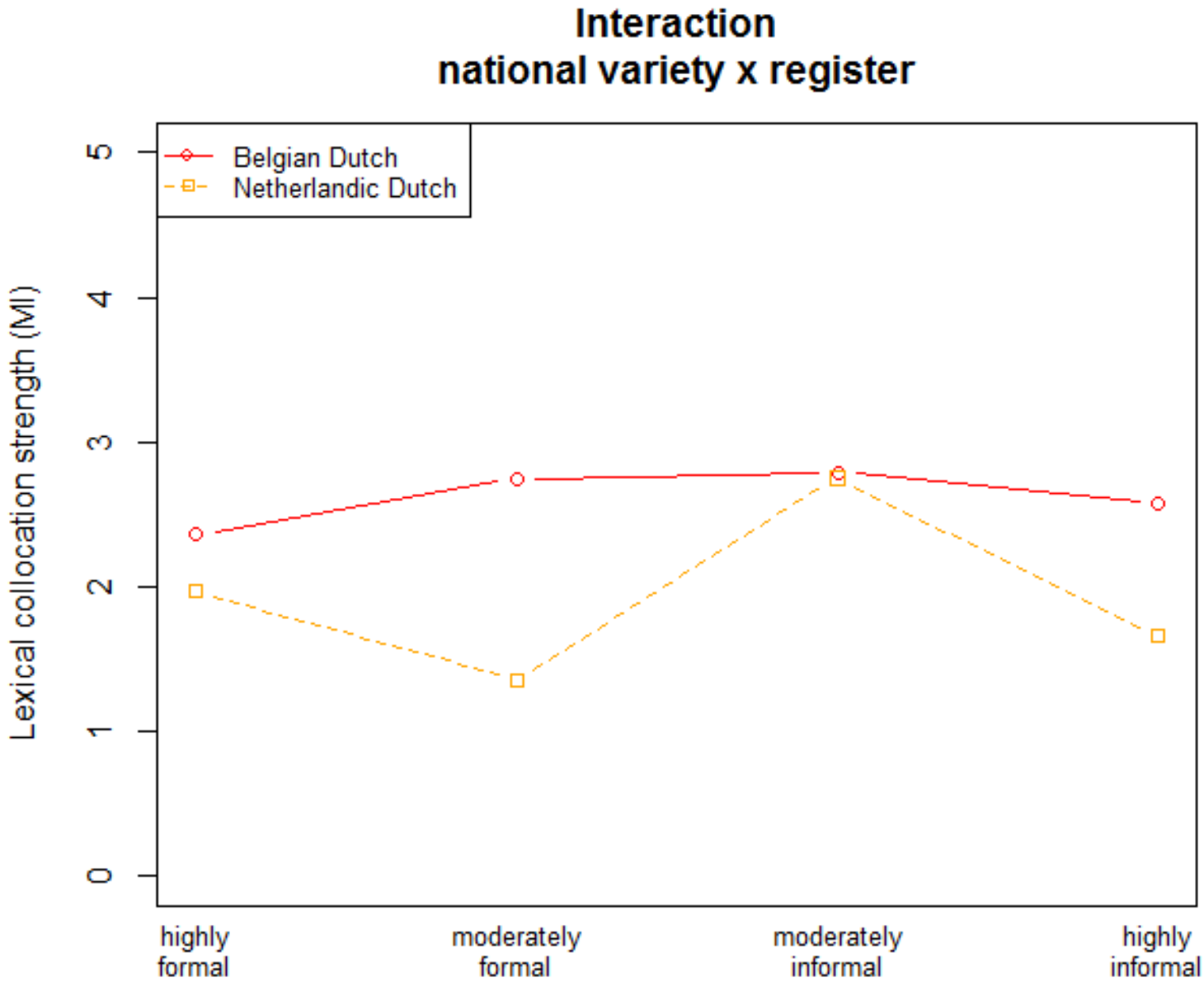
5. Results


Main effect national variety
Lexical collocation strength (MI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Belgian Dutch
Netherlandic Dutch
National variety

5. Results


Main effect register
Lexical collocation strength (MI)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
highly formal
moderately formal
moderately informal
highly informal
Register

5. Results


Interaction
national variety x register
Belgian Dutch
Netherlandic Dutch
Lexical collocation strength (MI)
0
1
2
3
4
5
highly formal
moderately formal
moderately informal
highly informal

5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Collocation strength AN pair

- Random effect (`speaker`):
  - Random intercept model: separate intercept fitted for each speaker
  - ICC = 0.12

5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

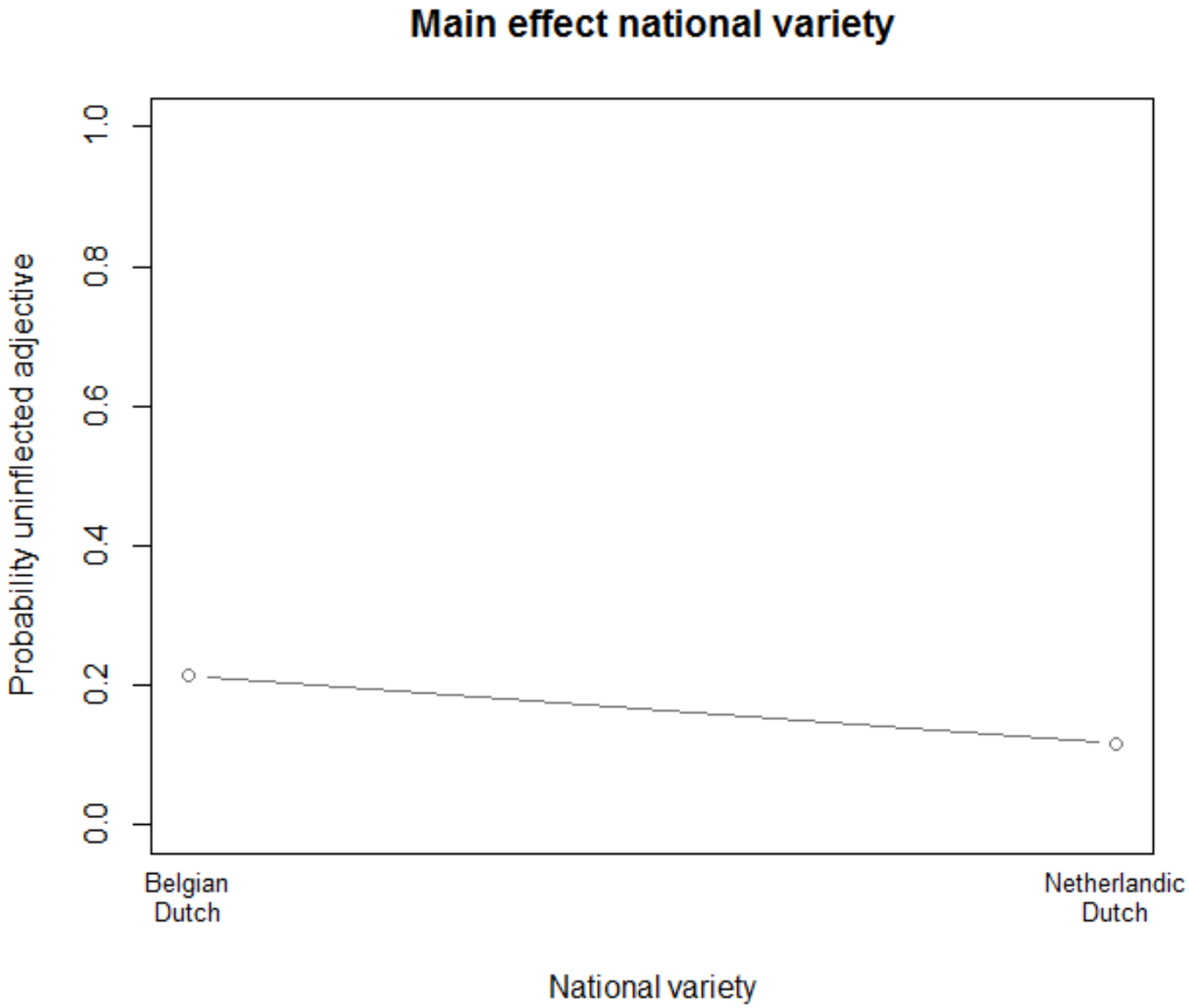
Adjectival inflectional alternation

- Model summary: sequential anova (Fox 2008)

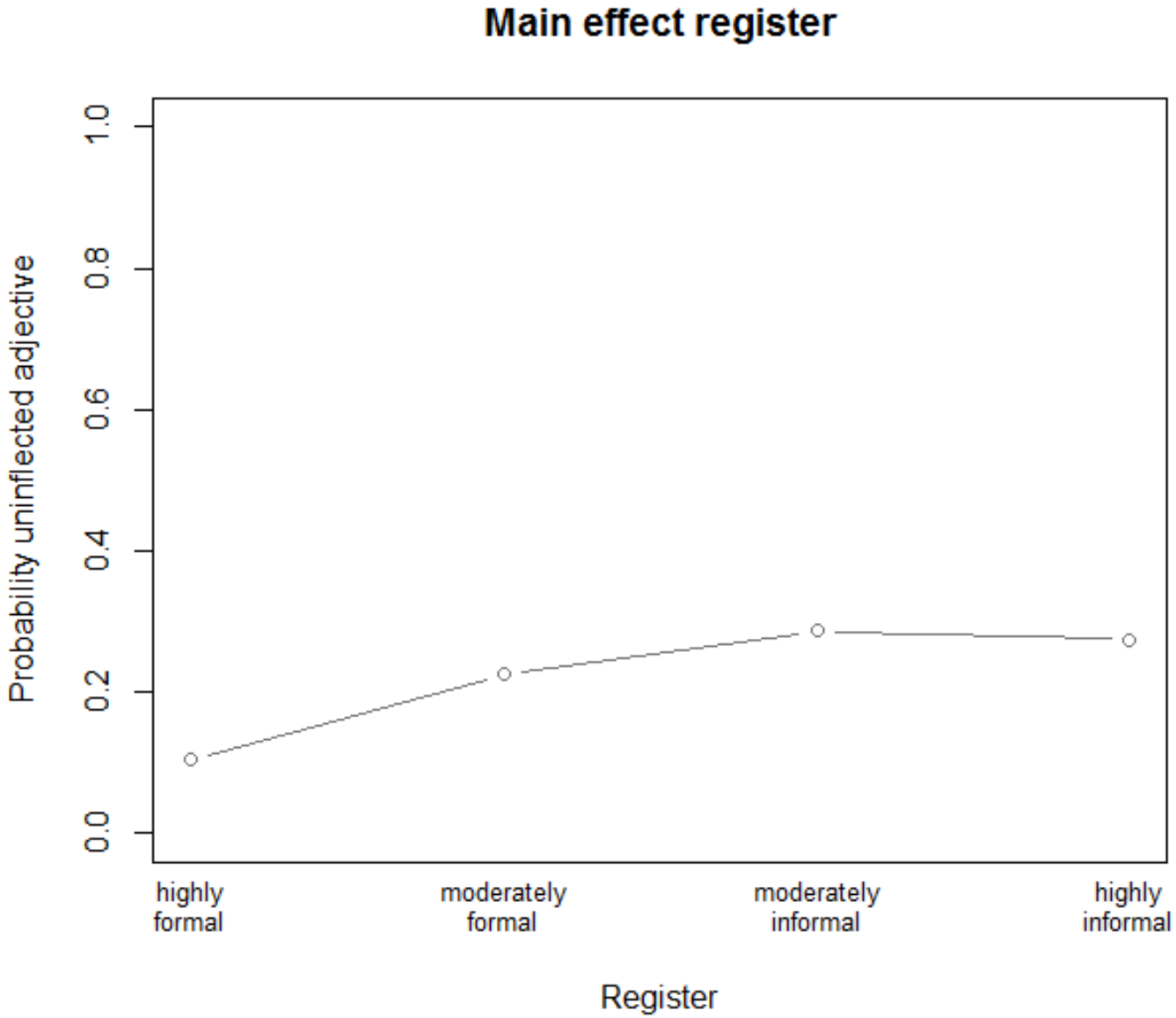
```
Analysis of Deviance Table (Type II Wald chisquare tests)

Response: infl
                          Chisq Df Pr(>Chisq)
nat.var                 40.9291  1  1.579e-10 ***
register               116.8310  3  < 2.2e-16 ***
lex.col                224.4876  1  < 2.2e-16 ***
nat.var:register        22.0001  3  6.523e-05 ***
nat.var:lec.col          0.6002  1    0.43851
register:lex.col        21.9796  3  6.587e-05 ***
nat.var:register:lex.col  7.2918  3    0.06316 .
```

- Overview fixed and random effects

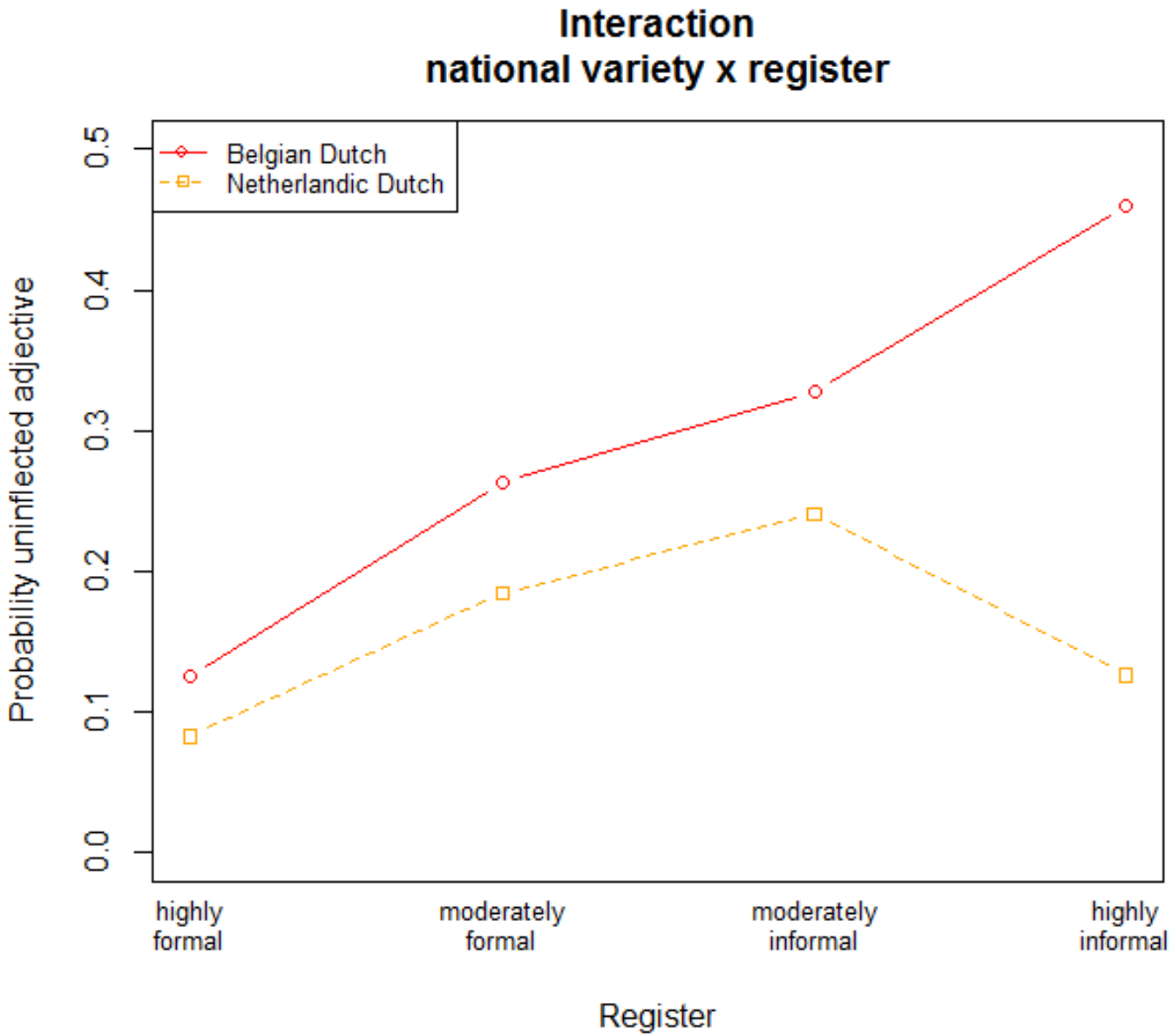
5. Results


Main effect national variety
Probability uninflected adjective
0.0
0.2
0.4
0.6
0.8
1.0
Belgian Dutch
Netherlandic Dutch
National variety

5. Results

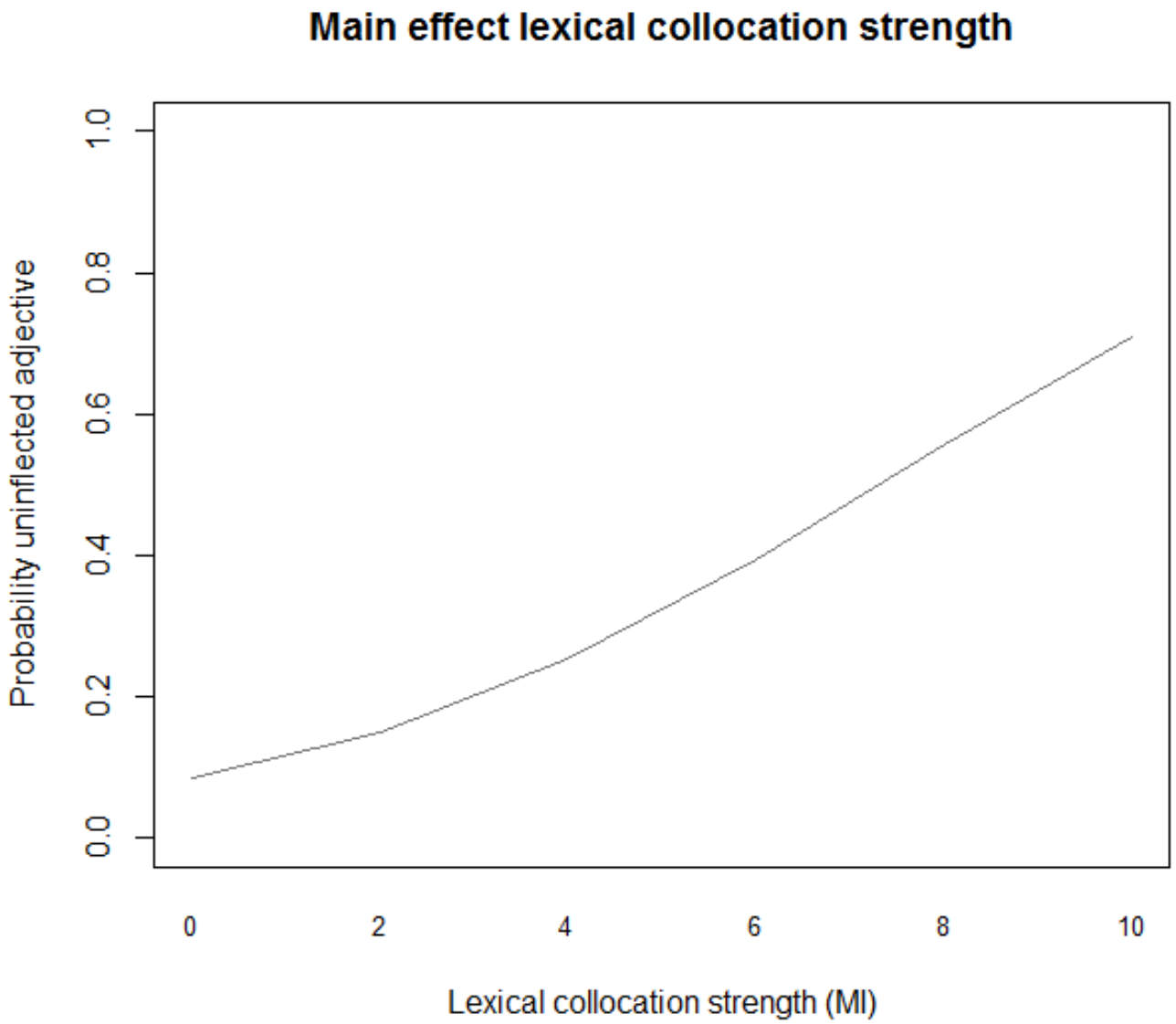

Main effect register
Probability uninflected adjective
0.0
0.2
0.4
0.6
0.8
1.0
highly formal
moderately formal
moderately informal
highly informal
Register

5. Results

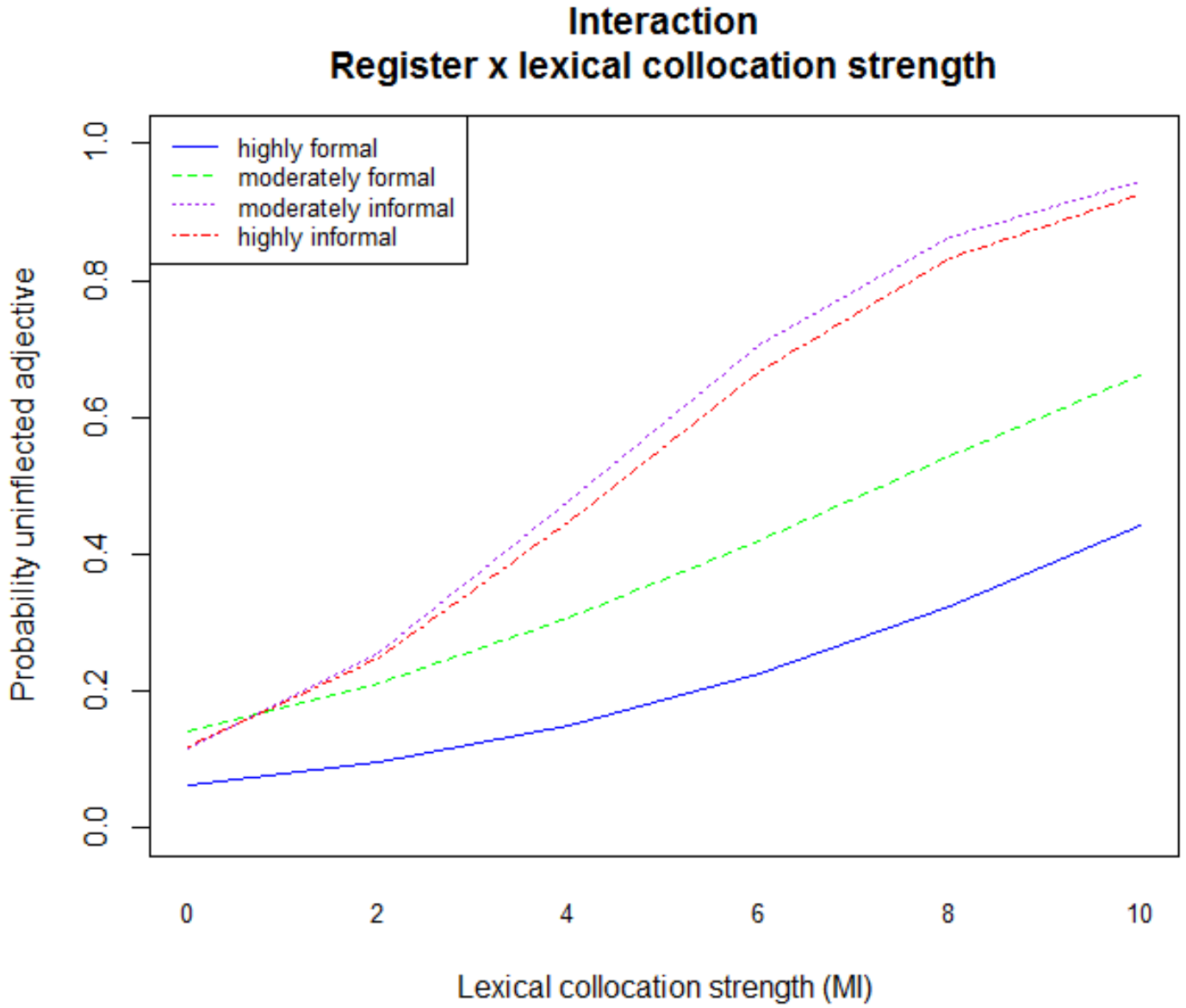

Interaction
national variety x register
Belgian Dutch
Netherlandic Dutch
Probability uninflected adjective
0.0
0.1
0.2
0.3
0.4
0.5
highly formal
moderately formal
moderately informal
highly informal
Register

5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN



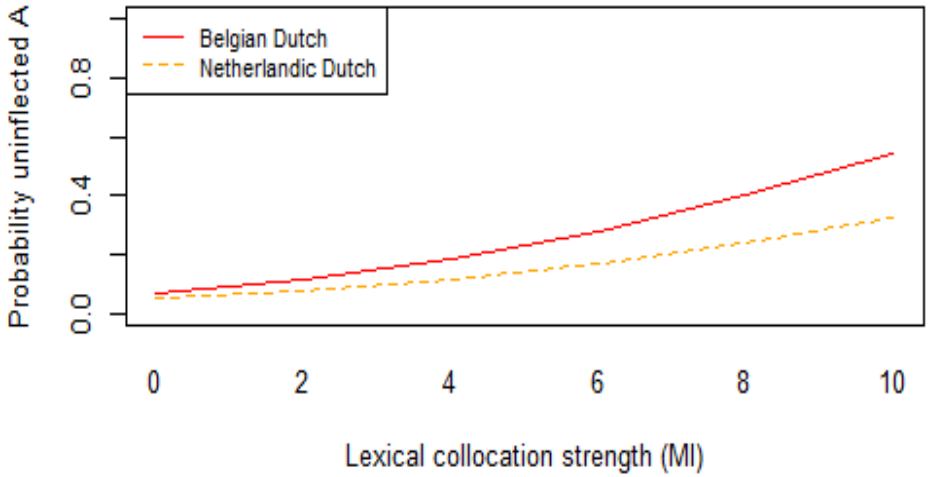
5. Results


Interaction
Register x lexical collocation strength
highly formal
moderately formal
moderately informal
highly informal
Probability uninflected adjective
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Lexical collocation strength (MI)

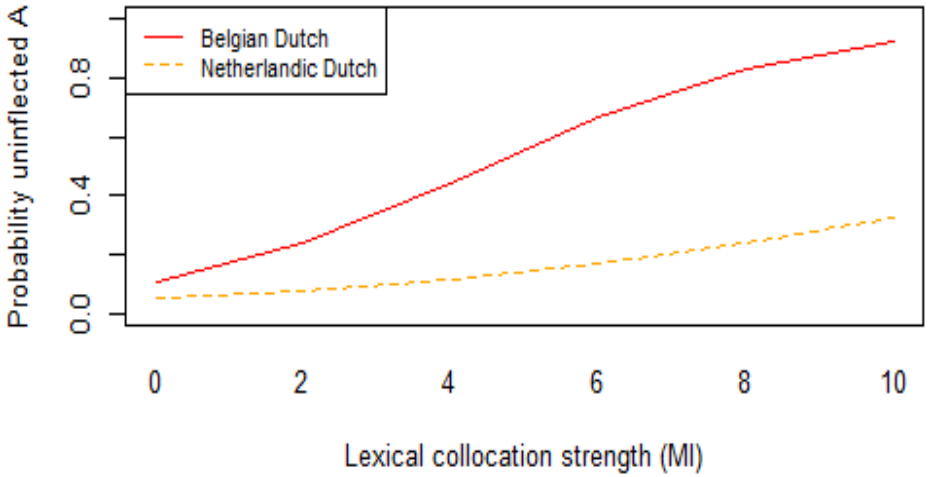
5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN

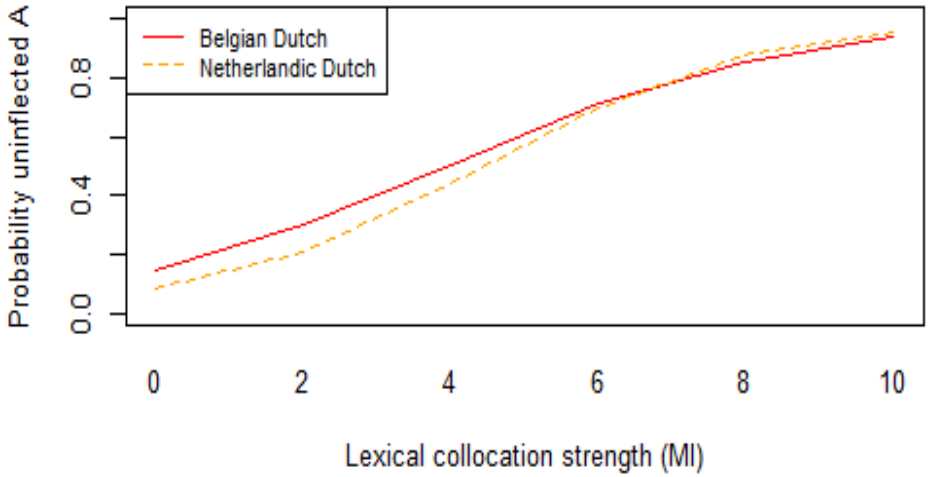
Highly formal register: Interaction lexical collocation strength x national variety



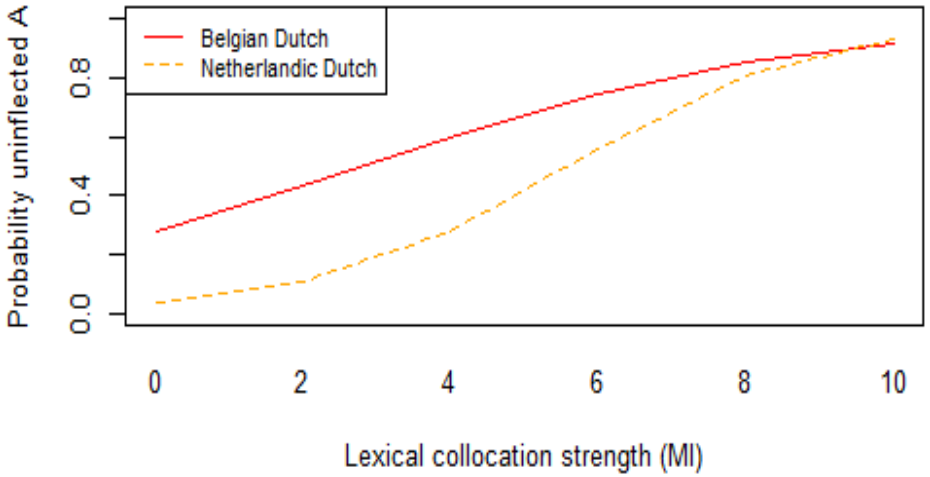
Moderately formal register: Interaction lexical collocation strength x national variety



Moderately informal register: Interaction lexical collocation strength x national variety

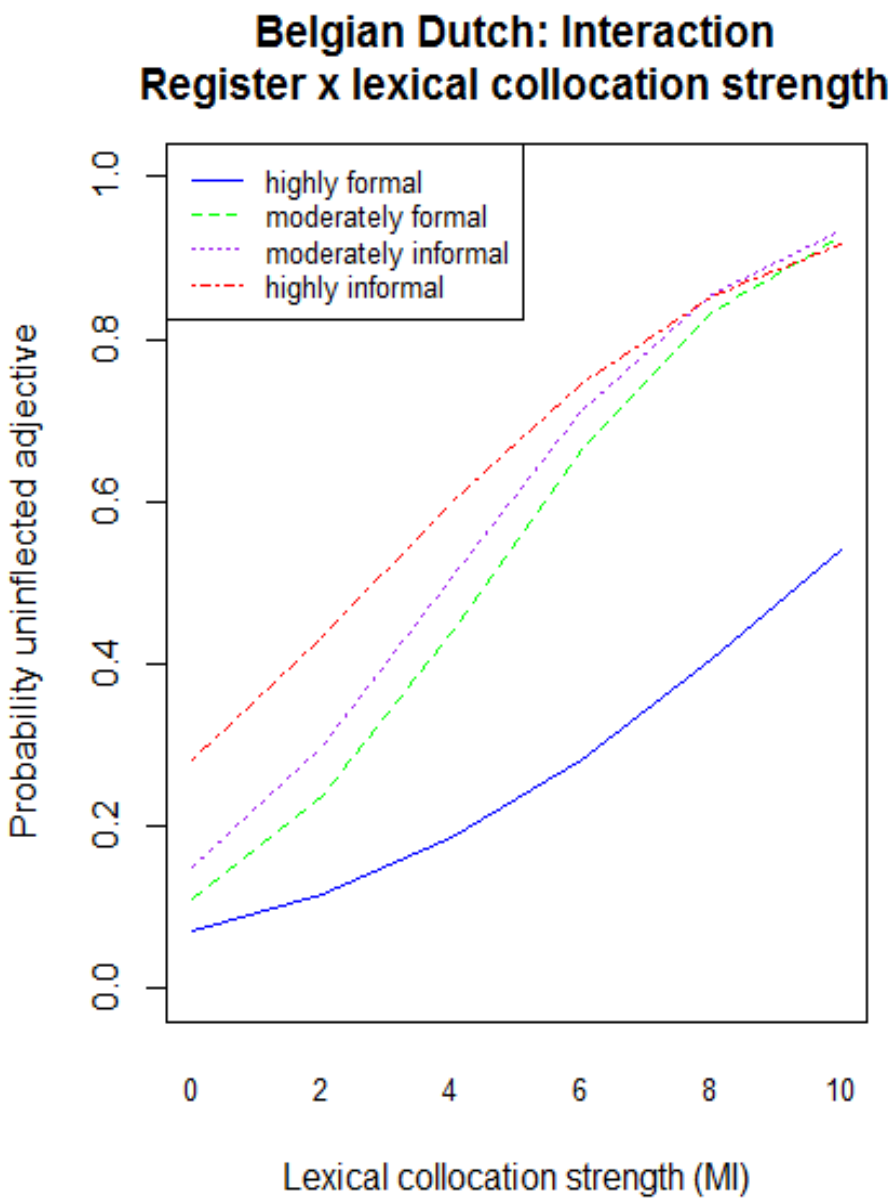


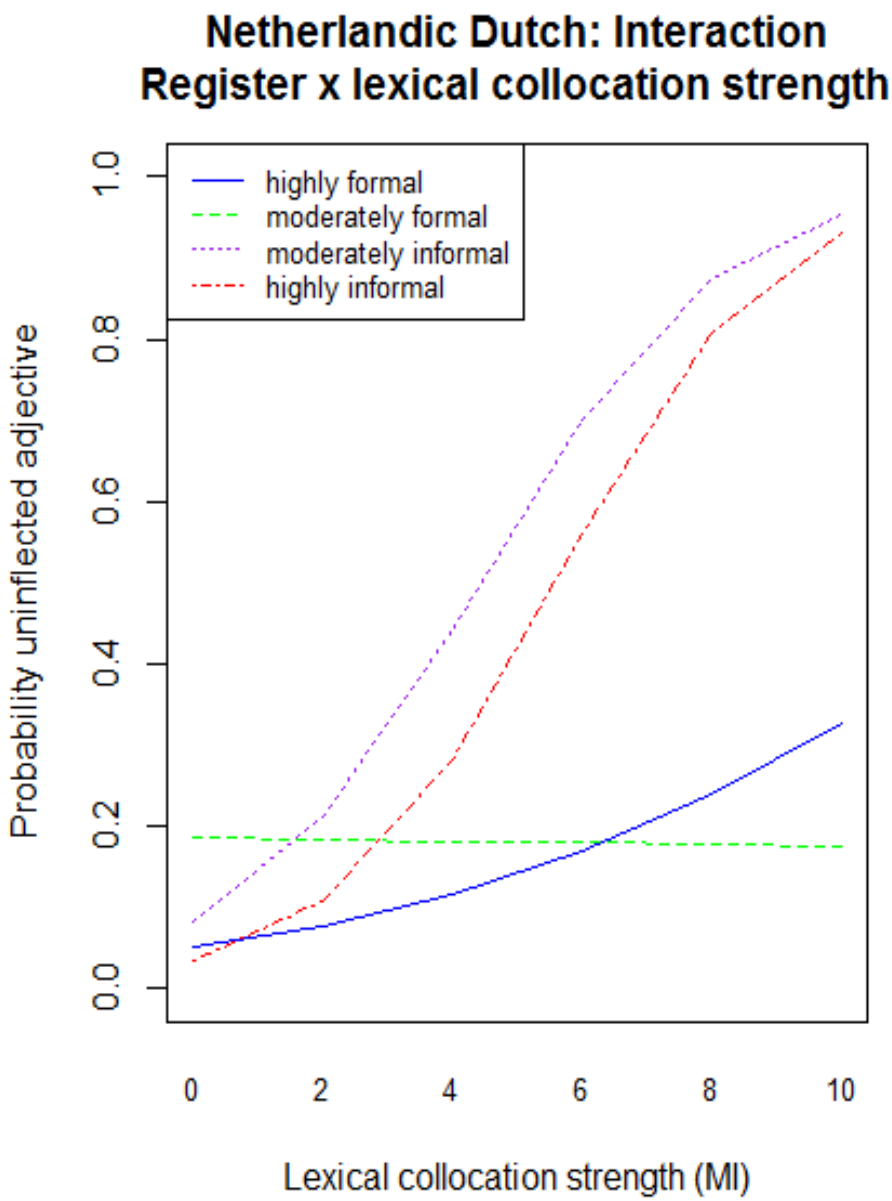
Highly informal register: Interaction lexical collocation strength x national variety



5. Results

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN



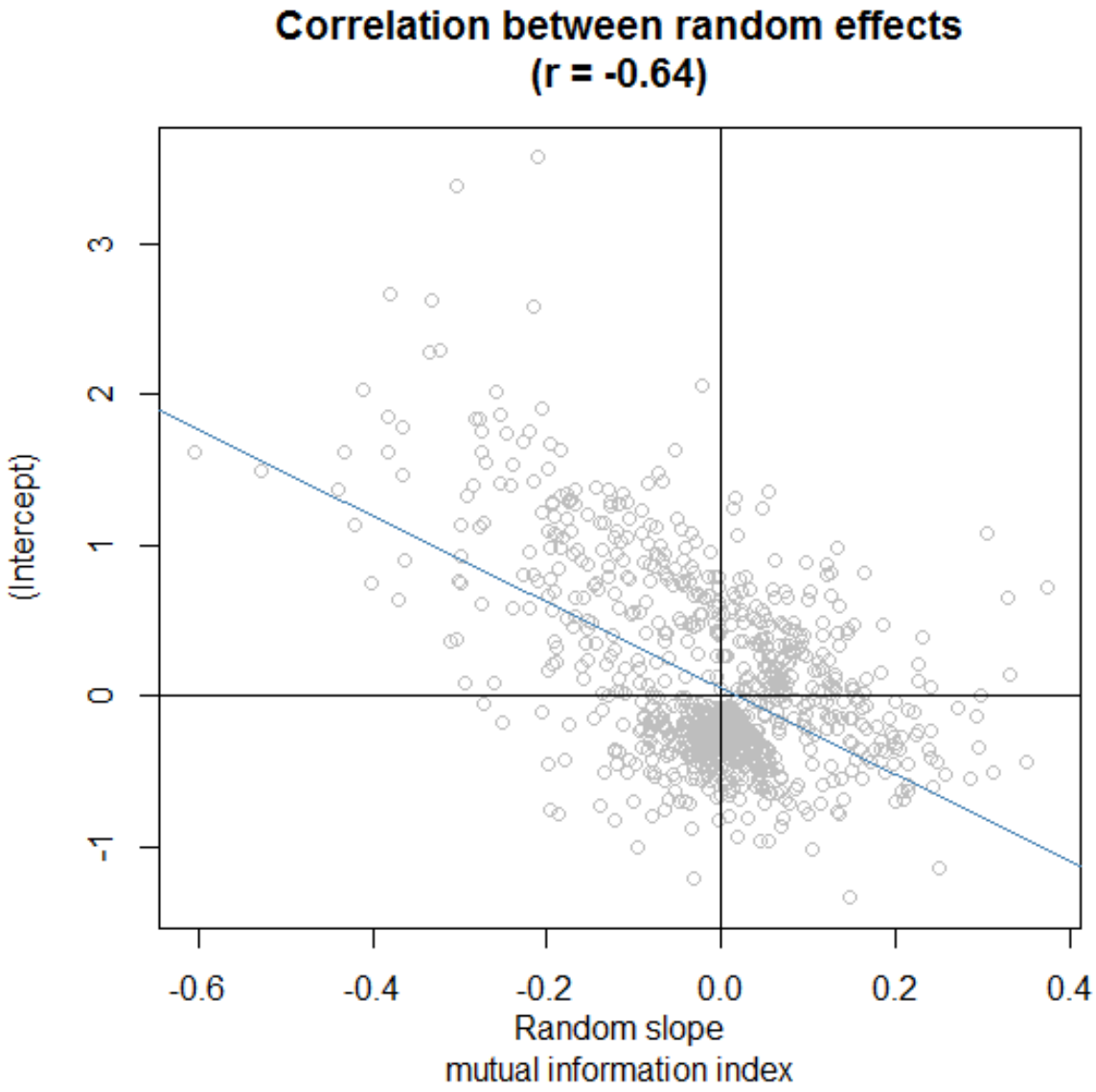


5. Results

Adjectival inflectional alternation

- Random effects:
  - Random intercept and random slope for `lex.col`
  - $ICC_{intercept}$ = 0.59
  - ICC_{slope} = 0.03
  - r(intercept,slope) = -0,64

5. Results


Correlation between random effects
(r = -0.64)
(Intercept)
3
2
1
0
-1
-0.6
-0.4
-0.2
0.0
0.2
0.4
Random slope
mutual information index

6. Discussion

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Results (1/2)

- Lexical collocation strength
  - No constant metric (as it is the case for word frequency; amongst others, Archer 2009; Baayen 2001; Brysbaert & New 2009)
  - Constrained by settings language use

- As lexical measure: constrained by
  - `nat.var`
  - `register`
  - `nat.var x register`
  - `speaker`'s idiosyncratic properties (cannot be reduced to `nat.var`)

- As determinant of adjectival inflection

6. Discussion

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Results (2/2)

- As determinant of adjectival inflection:
  - Main **deflecting effect**, mainly identifying
    - lexicalizing AN: categorizing adjectives, relational adjectives
    - lexicalized AN: institutional terms, proper names
  - Deflecting effect on adjectival inflection **constrained** by
    - `register`
    - `nat.var x register`
    - `speaker`'s idiolectic properties, where `lex.col` mainly compensates speakers with a low disposition toward uninflected adjective

6. Discussion

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Implications

- Usage settings cannot be discarded from corpus linguistic studies, since they affect basic corpus metrics

  - **Minimalist conception**: identification of usage settings to filter out potential constraints and biases induced by usage settings

  - **Maximalist conception:** full-fledged integration of settings of language use in corpus linguistic research

  (Geeraerts 2005)

Leuven University College, Marketing Communication

KU Leuven, Quantitative Lexicology and Variational Linguistics

jose.tummers@khleuven.be

dirk.speelman@arts.kuleuven.be

http://marco.khleuven.be

http://wwwling.arts.kuleuven.be/qlvl/