Lectal conditioning of lexical collocations

Jose Tummers
Dirk Speelman
Dirk Geeraerts

QITL-5 – Leuven, 12-14/09/2013

Contents

1. Problem statement
2. Case study
3. Data set
4. Research questions
5. Results
6. Discussion

1. Problem statement

Growing interest in authentic language data

- Probabilistic grammar
- Non-reductionist language models
- Language variation
- Corpus data

Linguistic frameworks

- Theoretical: Cognitive Linguistics, Construction Grammar
- Methodological: usage-based linguistics

3

1. Problem statement

Study of lexical preference patterns

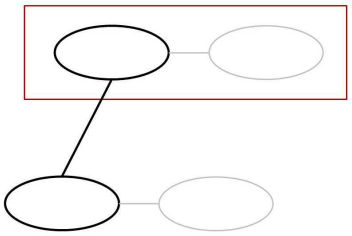
- Long-standing line of research in corpus linguistics: collocations and colligations (Firth 1957; Sinclair 1991; Hoey 1998)
- Alongside syntagmatic axis: collocations – relation between lexical items within construction (Speelman et al. 2009; Wulff 2008, 2013)
- Alongside paradigmatic axis: collostructions – relation between constructional slot and lexical instantiations (Stefanowitsch & Gries 2003, 2008)

4

1. Problem statement

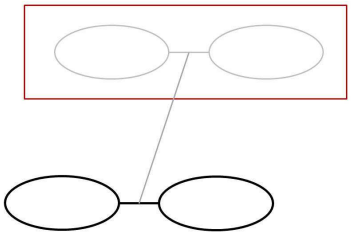
KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Collostructions



e.g. verbs associated to verbal slot in dative construction, such as *give*, *tell*, *send*, *offer*, *show* (Stefanowitsch & Gries 2003)

Collocations



e.g. strongly connected AN pairs, such as *openbaar vervoer* ('public transportation'), *vorig jaar* ('last year')

5

1. Problem statement

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Corpus data

- Representative sample of the language use of a given linguistic community in a/some given setting(s)
- Heterogeneous: (often) collected from different sources

Implication: socio-cultural diversity (Heylen et al. 2008)

- Analysis of **linguistic data**
- Analysis of properties of the **setting(s)**, including the heterogeneity of the linguistic community whose language use is represented = lectal dimension (Geeraerts 2005, 2013; Geeraerts et al. 2010)

6

1. Problem statement

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Lectal dimensions

- Sources of variation belonging to properties of settings of language use (language-external sources)
  - Dialect / regiolect / national variety
  - Sociolect
  - Register
  - …
- Caveat in mainstream (Cognitive) linguistics
  - Lectal variation analyzed in linguistic subdomains
  - Lexical patterning: frequency effects and levels of abstraction; phraseology
  - Exceptions:
    - Variational approaches (Grondelaers et al. 2008; Levshina et al. 2013; Szmrecsanyi 2010, 2013)
    - Collostructions (Stefanowitsch & Gries 2008)

7

2. Case study

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

Inflectional variation of adjective in Dutch NPs

- $[\text{DET}_{[+\text{DEFINITE}]}\ \text{ADJ}\ \text{N}_{[+\text{NEUTRAL},\ +\text{SINGULAR}]}]_{\text{NP}}$
- Adjectival inflection
  - -e (INFLECTED): unmarked and normative alternative
  - -ø (UNINFLECTED): marked alternative
- Example
  - *het vriendelijk-e kind*
    (the friendly-INFL child)
  - *het vriendelijk-ø kind*
    (the friendly-ZERO child)

8

2. Case study

Alternation governed by intricate network of variables
(Haeseryn et al. 1996; Rooij 1980; Tummers 2005)

- Structural
  - POS determiner, POS N
  - gradation A, gradation N
  - idiosyncrasy AN pair
- Lectal
  - national variety
  - register
- Discourse-processing
  - prosodic pattern AN pair
  - length A

9

2. Case study

Focus on

- Idiosyncrasy AN pair: uninflected adjective identifies AN as lexical unity (e.g. *kort geding* 'summary proceedings', *openbaar vervoer* 'public transportation')
- National variety: uninflected adjective is characteristic of Belgian Dutch as opposed to Netherlandic Dutch
- Register:
  - **Belgian Dutch**: uninflected adjective is characteristic of (highly) informal registers
  - **Netherlandic Dutch**: uninflected alternative is characteristic of highly formal registers (link with idiosyncrasy AN pair)

10

3. Data set

- *Corpus Gesproken Nederlands* (*Corpus of spoken Dutch*; Oostdijk 2001)

- Lectal organization of *Corpus of Spoken Dutch*
  - **National variety**: data realized by Netherlandic and Belgian Dutch speakers
  - **Register**: speech settings alongside 3 stylistic dimensions

FORMAL	INFORMAL
prepared	non-prepared
public	private
monologue	dialogue/multilogue

11

3. Data set

Response variable: adjectival alternation

	n	%
Inflected	3,810	76.75
Uninflected	1,154	23.25
Total	**4,964**	**1.00**

12

3. Data set

Operationalization explanatory variables

- **National variety** (`nat.var`):
  `Belgian.Dutch` vs. `Neth.Dutch`

- **Register** (`register`): based on 3 stylistic dimensions in corpus, 4 degrees of (in)formality are distinguished:
  `1high.form` > `2mod.form` > `3mod.inf` > `4high.inf`
  [0 1 2 3 informal stylistic values]
  e.g. dialogues = non-prepared & private & dialogue
  → `register = 4high.inf`

- **Lexical idiosyncrasy** (`llr`)
  - Lexical collocation strength
  - log likelihood ratio (G^2, Dunning 1993)
  - Measured between A and N lemmas

13

3. Data set

Operationalization explanatory variables

- Lexical idiosyncrasy (`llr`): **lexical collocation strength**
  - Qualitative criteria
    - "Fuzzy category" (Nunberg et al. 1994)
    - Conflicting syntactic test results (Matthews 1991)
    - Idiolectic differences (Moon 1998)
    - Prototypical instances of idiomatic expression
  - Quantitative measure
    - Gradual notion of idiomacy (Fillmore et al. 1988; Nunberg et al. 1994)
    - Continuum ranging
      from fixed lexical sequences (e.g. *half uur* 'half hour')
      over formulaic expressions (e.g. *geregistreerd partnerschap* 'registered partnership')
      to naming expressions (e.g. *Vlaams Parlement* 'Flemish Parliament')

14

4. Research questions

1. Is the lexical collocation strength (`llr`) lectally constrained?

2. Is the impact of the lexical collocation strength (`llr`) on the adjectival inflection lectally constrained?

15

4. Research questions

1. **Is the lexical collocation strength (`llr`) lectally constrained?**

2. Is the impact of the lexical collocation strength (`llr`) on the adjectival inflection lectally constrained?

16

5. Results : Research Question 1

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

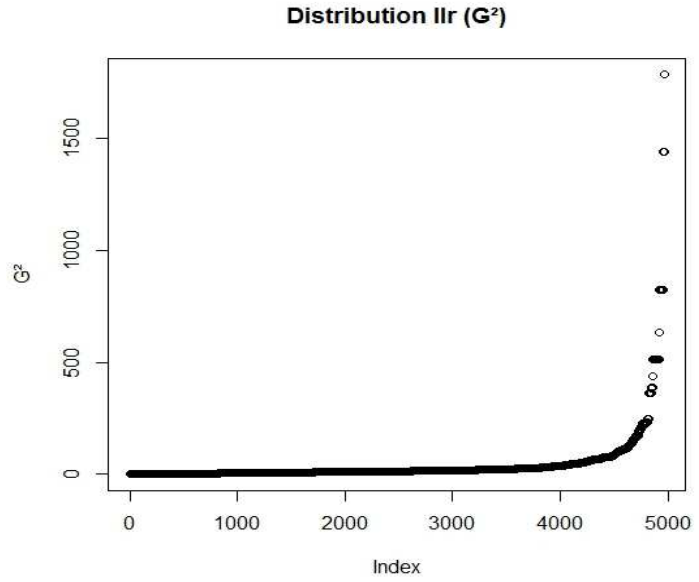
Log likelihood ratio (`llr`):

- heavily biased distribution
- G^2 = -2 log likelihood ratio : X^2 distribution (Dunning 1993) which is a specific subtype of the gamma distribution (Forbes et al. 2011)
- `glm`
  - `family = gamma`
  - `link = inverse`
  - `llr ~ nat.var * register`

17

5. Results : Research Question 1

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN



18

5. Results : Research Question 1

Model statistics

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.4504  -1.4085  -0.8242  -0.2159   9.4580

Null deviance:     14564  on 4963  degrees of freedom
Residual deviance: 13319  on 4956  degrees of freedom
```

Model significance

```
> 1 - pchisq(14564 - 13319, 4963 - 4956)
[1] 0
```

19

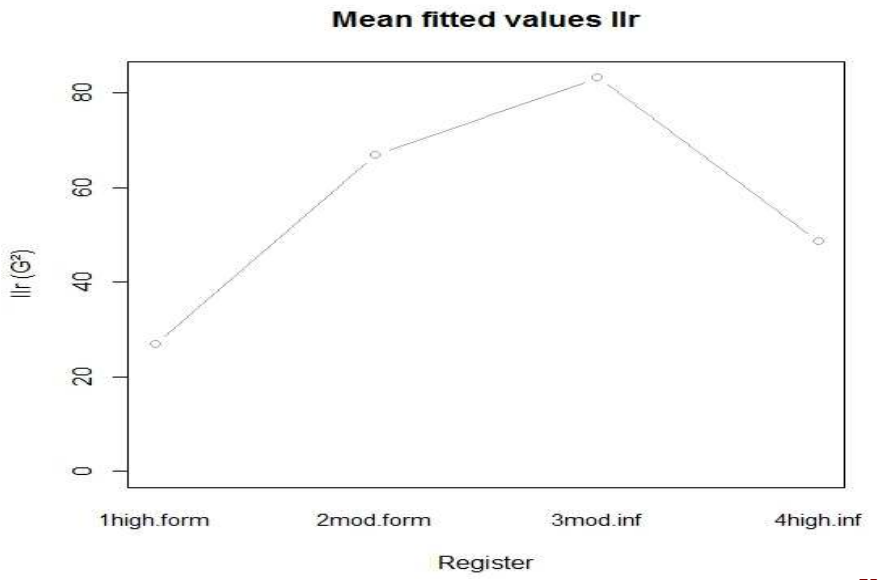
5. Results : Research Question 1

Coefficients

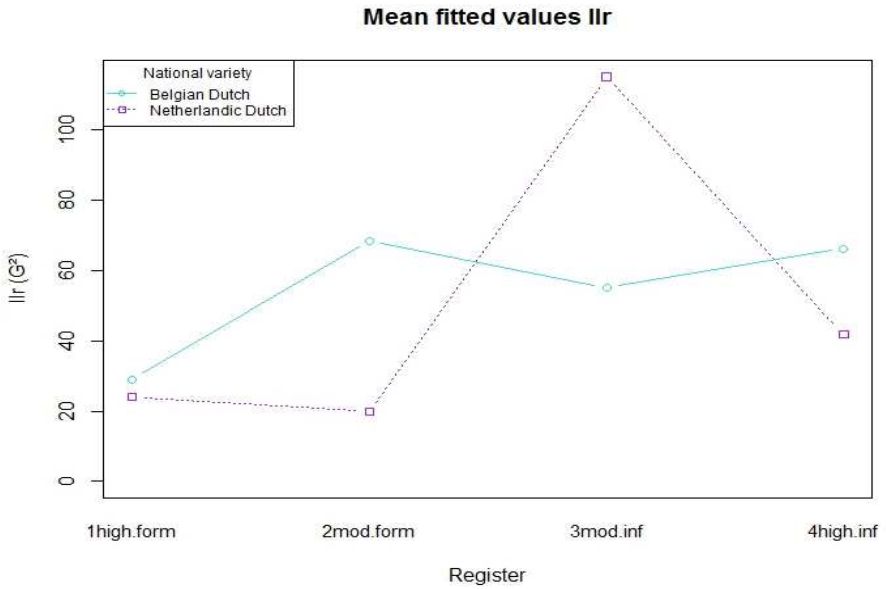
	Est.	SE	t value	Pr(>|t|)	
(Intercept)	0.034	0.002	16.172	< 2e-16	***
nat.var=Neth.Dutch	0.007	0.003	1.907	0.056541	.
register=2mod.form	**-0.019**	**0.002**	**-7.206**	**6.61e-13**	***
register=3mod.inf	**-0.016**	**0.003**	**-5.323**	**1.06e-07**	***
register=4high.inf	**-0.019**	**0.003**	**-6.365**	**2.13e-10**	***
nat.var=Neth.Dutch:register=2mod.form	0.028	0.032	0.896	0.370511	
nat.var=Neth.Dutch:register=3mod.inf	**-0.016**	**0.004**	**-3.730**	**0.000194**	***
nat.var=Neth.Dutch:register=4high.inf	0.001	0.004	0.375	0.708046	

20

5. Results : Research Question 1



5. Results : Research Question 1



5. Results : Research Question 1

Summary: Lectal constraining of lexical collocation strength

- No significant main effect `nat.var`
- Significant main effect `register`: `llr` ~ informality
- Interaction: high mean `llr` in `nat.var=Neth.Dutch` & `register=3mod.inf`
  - `register=3mod.inf`: dialogue/multilogue & spontaneous & public
  - **Topical bias**:
    - Debates in Dutch parliament: 355/366 observations
    - Overrepresentation of highly formulaic administrative language use

23

4. Research questions

1. Is the lexical collocation strength (`llr`) lectally constrained?

2. **Is the impact of the lexical collocation strength (`llr`) on the adjectival inflection lectally constrained?**

24

5. Results : Research Question 2

Lectally constrained impact of `llr` on inflectional alternation

- Possible outcomes
  - Adjectival inflection is lectally conditioned
    - ☞ significant impact `nat.var` and/or `register`, no significant impact `llr`
  - Adjectival inflection is lexically conditioned by lexical collocations
    - ☞ significant impact `llr`, no significant impact `nat.var` nor `register`
  - `llr` and lectal variables independently condition inflectional alternation
    - ☞ significant main effects `llr` and `nat.var` & `register` without significant interaction between `llr` and lectal variables
  - Impact `llr` is lectally constrained
    - ☞ significant interaction(s) between `llr` and `nat.var` & `register`

25

5. Results : Research Question 2

- Logistic regression analysis (`rms` library; Harrel 2001)

$$\log\left(\frac{a.uninflected}{a.inflected}\right) \sim llr * nat.var * register$$

  - **Positive** coefficient: variable value favoring **uninflected A** compared to reference value
  - **Negative** coefficient: variable value favoring **inflected A** compared to reference value

- Model statistics

$LR\ Chi^2 = 648{,}11,\ df = 15,\ p < 0.0001$

$C = 0.732$

26

5. Results : Research Question 2

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN

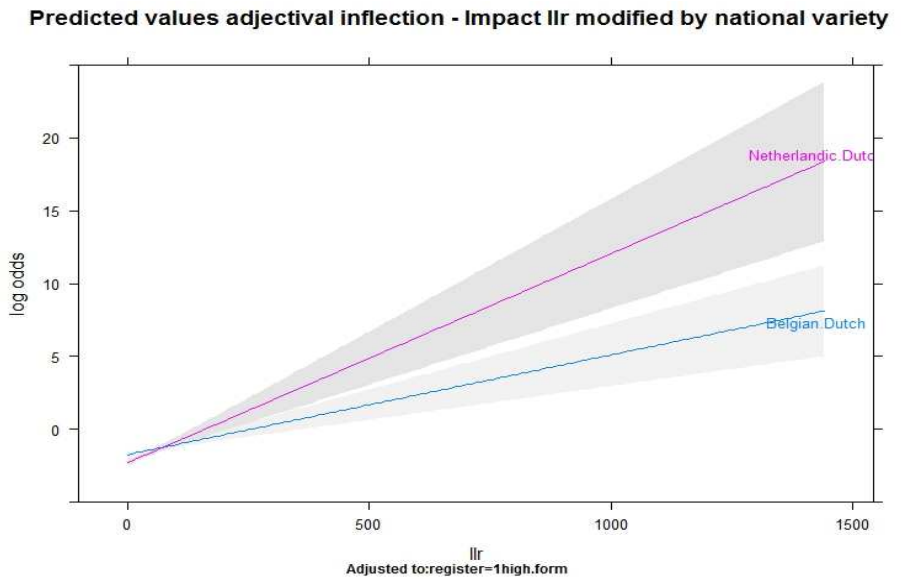
Coefficients

	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	-1.7572	0.0768	-22.88	<0.0001
llr	**0.0069**	**0.0011**	**6.06**	**<0.0001**
nat.var=Neth.Dutch	**-0.5516**	**0.1352**	**-4.08**	**<0.0001**
register=2mod.form	**0.7489**	**0.1464**	**5.11**	**<0.0001**
register=3mod.inf	**1.3220**	**0.1350**	**9.79**	**<0.0001**
register=4high.inf	**1.4318**	**0.1537**	**9.32**	**<0.0001**
nat.var=Neth.Dutch * register=2mod.form	1.4334	1.0215	1.40	0.1605
nat.var=Neth.Dutch * register=3mod.inf	0.0682	0.2170	0.31	0.7532
nat.var=Neth.Dutch * register=4high.inf	**-1.2169**	**0.2277**	**-5.34**	**<0.0001**
llr * nat.var=Neth.Dutch	**0.0075**	**0.0023**	**3.29**	**0.0010**
llr * register=2mod.form	-0.0015	0.0015	-0.98	0.3258
llr * register=3mod.inf	**-0.0047**	**0.0014**	**-3.26**	**0.0011**
llr * register=4high.inf	-0.0033	0.0019	-1.79	0.0728
llr * nat.var=Neth.Dutch * register=2mod.form	-0.0627	0.0509	-1.23	0.2178
llr * nat.var=Neth.Dutch * register=3mod.inf	**-0.0054**	**0.0026**	**-2.11**	**0.0350**
llr * nat.var=Neth.Dutch * register=4high.inf	0.0001	0.0032	0.02	0.9811

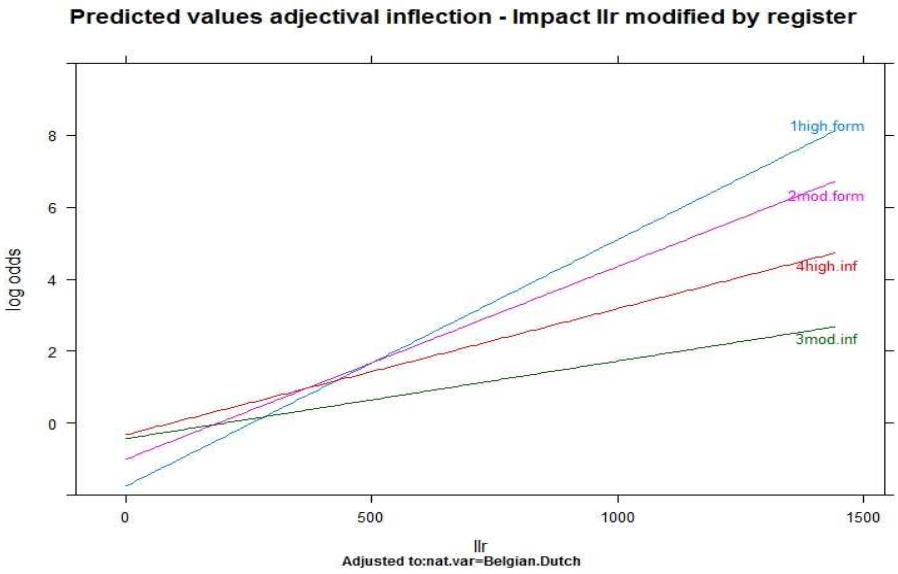
27

5. Results : Research Question 2

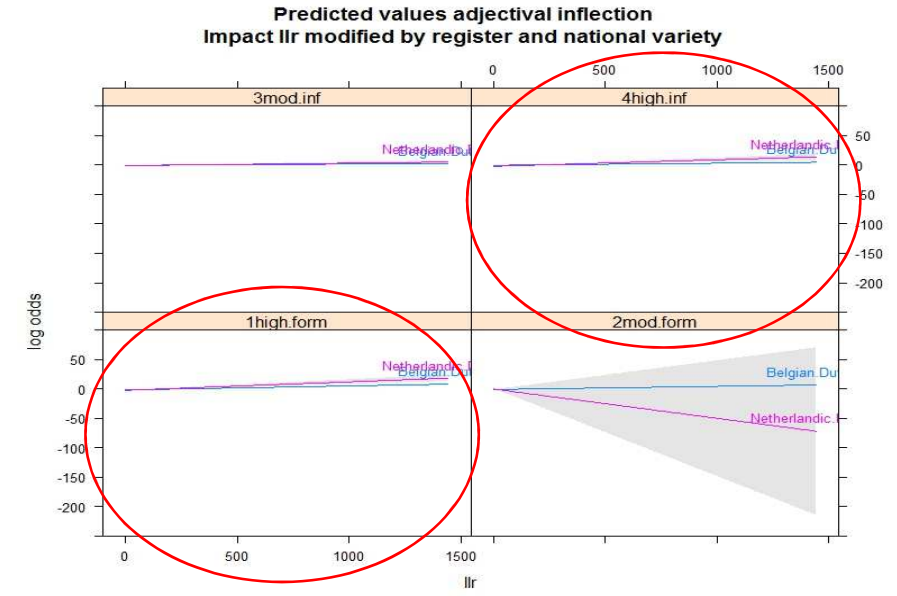
KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN



5. Results : Research Question 2

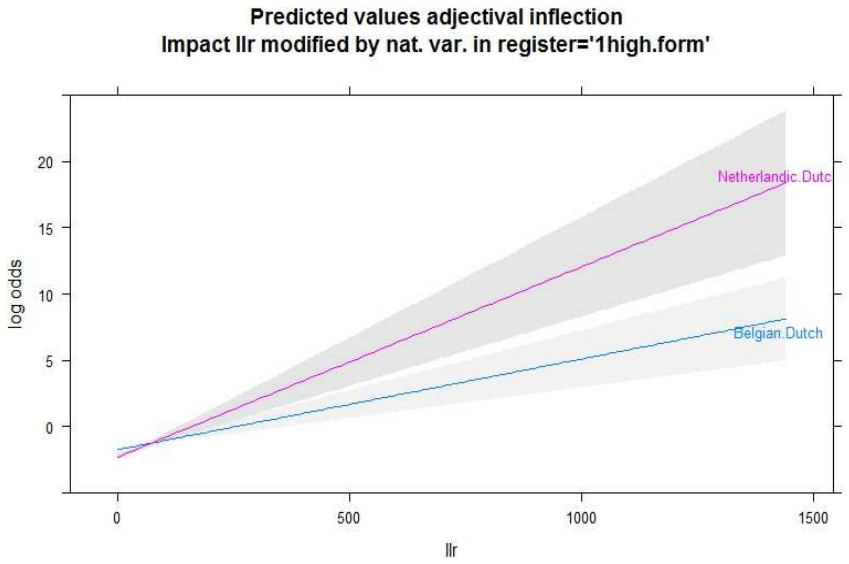


5. Results : Research Question 2



5. Results : Research Question 2

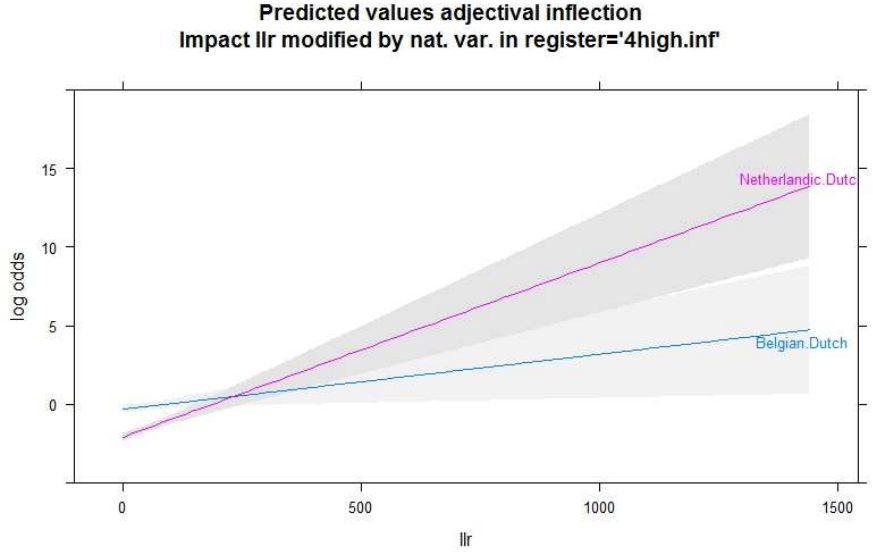
KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN



31

5. Results : Research Question 2

KHLEUVEN
KATHOLIEKE HOGESCHOOL LEUVEN
ASSOCIATIE KATHOLIEKE UNIVERSITEIT LEUVEN



32

5. Results : Research Question 2

Summary (1/2)

- Lectal dimensions
  - `nat.var`: tendency to use uninflected A in Belgian Dutch
  - `register`: tendency to use uninflected A in informal registers
  - `nat.var * register`: stronger tendency to use uninflected A in (highly) informal registers in Belgian Dutch
- Lexical collocation strength (`llr`): positive impact on selection uninflected A
- Lectal constraints on impact collocation strength

33

5. Results : Research Question 2

Summary (2/2)

- Lectal constraints on impact collocation strength (`llr`)
  - `nat.var`: impact `llr` on selection uninflected A higher in `Netherlandic.Dutch`
  - `register`: impact `llr` on selection uninflected A lower in `3mod.inf`
  - `nat.var * register`: impact `llr` on selection uninflected A lower in `3mod.inf` in `Netherlandic.Dutch`

34

6. Discussion

Lexical collocations (`llr`): lectally constrained

- Collocation strength
  - `register`: Further research to disentangle register components (topic, speakers' ID, medium, etc.)
  - `register * nat.var`: corpus-specific restrictions
- Impact `llr` on adjectival inflection
  - `register`
  - `nat.var`
  - `nat.var * register`
- Lexical collocation strength should be measured taking into account the lectal structure of the corpus

35

6. Discussion

Adjectival inflection (`llr`):

- Determinants of uninflected A
  - Lexical collocation strength AN pair
  - Lectal variables
    - National variety: Belgian Dutch
    - Register: informal registers
    - Interaction
- Interactions triggering use of uninflected A
  - Netherlandic Dutch: lexical collocations and formulaic language
  - Belgian Dutch:
    - Lexical collocations and formulaic language (exogenous use)
    - Informal registers (endogenous use)

36

6. Discussion

Implications for usage-based linguistic theory (1/2)

- **Impact of settings language use** in a usage-based grammar
  - Constructional constraining
    e.g. impact `register` and `nat.var` on inflectional alternation
    e.g. impact of `register` on `llr`
  - Variable constraining
    e.g. altered impact of `llr` on inflectional variation according to `register`, `nat.var` as well as their interaction
  - Lectal dimension interacting with structural and processing dimensions of usage-based grammar (Levshina 2013; Stefanowitsch & Gries 2008)

37

6. Discussion

Implications for usage-based linguistic theory (2/2)

- **Settings of language** use – as present in corpus design – have to be included in usage-based language models
- **Dimensions of meaning** in usage-based grammar (Geeraerts et al. 2010; Kristiansen 2006)
  - Conceptual meaning (~ ideational function)
  - Lectal/social meaning (~ interpersonal function)
- Language as "**diasystem**" (Geeraerts 2005; Geeraerts et al. 2010)
  - System of overlapping language repertoires, activated according to usage settings
  - Cognitive Sociolinguistics

38

6. Discussion

Linguistic varieties vs. different languages

- Language **system**: How much difference between language varieties can be borne by a linguistic model, even in a diasystem?
- Language **use**: competition between expression (social/lectal meaning) and intelligibility
- **Criteria**
  - Perceptual studies (Grondelaers et al. 2013)
  - Mutual intelligibility (Impe 2011)

39

Leuven University College
KULeuven, Quantitative Lexicology and Variation Linguistics

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

jose.tummers@khleuven.be
dirk.speelman@arts.kuleuven.be
dirk.geeraerts@arts.kuleuven.be

40