

A Lemmatized Concordance of the Works of *Jan van Ruusbroec*

Guy De Pauw

Mike Kestemont, Walter Daelemans,
Thom Mertens, Guido De Baere

GALATEA II – Antwerp, Belgium
18/2/2011



The Works of Jan Van Ruusbroec

- Flemish mystic
- 1294 – 1381
- John of Ruysbroeck



	# tokens	index
Dat rijcke der ghelieven	30 689	+
Die geestelike brulocht	46 607	+
Vanden blinkenden steen	11 901	+
Vanden vier becoringhen	4 307	+
Vanden kerstenen ghelove	5 386	+
Vanden geesteliken tabernakel	128 585	±
Vanden seven sloten	13 666	-
Van seven trappen	14 964	+
Een spiegel der ewigher salicheit	28 608	+
Boecksken der verclaringhe	7 004	-
Vanden XII beghinen	70 220	+
Brieven (7 + 1)	5 677	+
Total	367 614	

Initial Wish List

- Ruusbroecgenootschap 2008
http://www.ua.ac.be/main.aspx?c=*RUUSBROEC
- Lemmatized Concordance of the works of Jan van Ruusbroec
 - Concordance of word forms
 - Lemmatized concordance in order of appearance
 - Lemmatized concordance in order of word forms
 - Lemmatized concordance in order of right context
 - Part-of-speech tags for ambiguous words (e.g. *gevallen* v/n)
 - **Digital:** Complete concordance
 - **Hardcopy:** Only selection

Bottlenecks

- Which lemma?
- Norm: MNW?
- What to do with separable verbs?
in den vs inden
- Part-of-speech tagging
beminde participle or noun?
- Names and Latin words
- Words not in MNW?

Where to start?

- Spelling variation
- No digital copy of MNW
- Find lookalike word forms using a maximum entropy classifier

De Pauw, G., Wagacha P. W. & D.A. Abade (2007) Unsupervised Induction of Dholuo Word Classes using Maximum Entropy Learning. Proceedings of the First International Computer Science and ICT Conference. University of Nairobi: Nairobi, Kenya.

andiko	I=n I=ndi E=iko B=an I=k I=d E=o I=di E=ndiko B=andik B=a I=ik I=nd I=ndik I=i E=ko
gindiko	I=n I=ndi B=g E=iko B=gin I=k I=d B=gi E=o I=di E=ndiko I=in I=ik I=nd I=ndik I=i I=indik E=ko

Lookalike-concordance

- Proof-of-the-principle on-line demo
- No lemma, just overlapping “groups” of lookalikes
- Because of extensive spelling variation, limited morphology and high-rate of OOVs, quite a lot of false positives and true negatives

e.g. ommegaen => gaen
 ommevaen => varuwen
 wesen => menschen

Not usable for target audience, but interesting, faster alternative to edit distance metrics

- Trip to INL, Leiden: digital copy of MNW

```
<title>ACHTAGE</title>
<index><lemma>ACHTAGE</lemma></index><nrind>00052</nrind></header>
<header>
<title>ACHTBAER</title>
<index><lemma>ACHTBAER</lemma></index><nrind>00053</nrind></header>
<header>
<title>ACHTE</title>
<index><lemma>ACHTE</lemma></index><nrind>00054</nrind></header>
<header>
<title>ACHTE</title>
<index><lemma>ACHTE</lemma></index><nrind>00055</nrind></header>
<header>
<title>ACHTE</title>
<index><lemma>ACHTE</lemma></index><nrind>00056</nrind></header>
<header>
<title>ACHTE</title>
<index><lemma>ACHTE</lemma></index><nrind>00057</nrind></header>
<header>
<title>ACHTE</title>
<index><lemma>ACHTE</lemma></index><nrind>00058</nrind></header>
<header>
<title>ACHTE</title>
<index><lemma>ACHTE</lemma></index><nrind>00059</nrind></header>
<header>
<title>ACHTEEL</title>
<index><lemma>ACHTEEL</lemma></index><nrind>00060</nrind></header>
```



- Match word forms in Ruusbroec data to the “closest” entry in the MNW lexicon
- Problems:
 - Which one to pick?
 - Match to word forms or lemmas?
 - Different metrics tried: Levenshtein, Dice, maxent-based
 - No gold-standard, difficult to evaluate, tweak & tune
 - Lemmatize function words, content words?
 - Ambiguity: ingel => engel, hengel
 - OOVs

Plan B

- Proof-of-the-principle on-line demo
- Comments:
 - Too many mistakes to consider this as a pre-processing step for manual correction
 - Too many OOVs

Not usable for target audience, but good start

But: new information source surfaced!

Indexes

	# tokens	index
Dat rijcke der ghelieven	30 689	+
Die geestelike brulocht	46 607	+
Vanden blinkenden steen	11 901	+
Vanden vier becoringhen	4 307	+
Vanden kerstenen ghelove	5 386	+
Vanden geesteliken tabernakel	128 585	±
Vanden seven sloten	13 666	-
Van seven trappen	14 964	+
Een spiegel der ewigher salicheit	28 608	+
Boecksken der verclaringhe	7 004	-
Vanden XII beghinen	70 220	+
Brieven (7 + 1)	5 677	+
Total	367 614	

```

<fi,macro>
<voetn:$(((voetnoot$: &1&)))$>
<s,2><cb>VOCABULARY<ce>
<rh,$(VANDEN VIER BECORINGHEN - VOCABULARY<st><dpn>$)>
<rf, >-
<np>
<s,2> For the introduction to the vocabulary see <ub>Vanden blinkenden
steen<ue>, pp. 000-000.<bl>1<br> The basic MS for the text of
<ub>Vanden vier becoringhen<ue> is the MS F (Brussels, Koninklijke
Bibliotheek, 1165-67).
<s,25> <ub>Voetnoot bij de inleiding<ue>
<s,1> (1) For the <ub>List of the editor's corrections in the text of
MS F<ue> see pp. 000-000.--- The following words do not occur in the
<ub>Middelnederlandsch Woordenboek<ue>: <ub>abstinencie<ue>,
<ub>adel<ue>, <ub>affectie<ue>, <ub>afgod<ue>, <ub>apostel<ue>,
<ub>doregronden<ue>, <ub>jubileus<ue>, <ub>ongeachtsam<ue>,
<ub>overweselijcheit<ue>.-
<np>
<hi,3>%A-
<nl>-
<s,2>%abijt: clothing, n. 67, 99-
<nl>*abstinencie: abstinence, n. 231-
<nl>%achten: esteem, think, v. 120, 163, 176, 297-
<nl>*adel: nobility, n. 331-
<nl>%aengripen: take up, v. 99, *124, 221-
<nl>%aencleven: cling to, v. 207-
<nl>%aenroepen: invoke, v. 262, 268, 274, 280-
<nl>%aenscijn: countenance, face, n. 18, 173, 241-
<nl>*affectie: affection, n. 295-
<nl>%affectioes: affectional, adj. 299-
<nl>*afgod: idol, n. 209-
<nl>%afgrondicheit: fathomlessness, n. 280-
<nl>%afnemen: wane, v. 306-
<nl>%almeest: for the greater part, adv. 172
<nl>%almuegent: omnipotent, adj. 258-
<nl>%alremeest: most of all, adj. 60
<nl>%alremeest: most of all, adv. 181
<nl>%altoos: always, adv. 14, 15, 16, 21, 36, 53, 128, 168, 220, 335-
<nl>*apostel: apostle, n. 27-
<nl>%arbeiden: work, v. 304, 314-
<nl>%arbeit: work, n. 324-

```

Indexes

**7 different
formats (!)**

Plan C

We now know which lemmas occur on which lines!

- Convert indices to database
- Calculate for each line:
 - Levenshtein distance between
 - Each word form
 - Each lemma in this line (according to index)
 - Match from low->high distances
- Proof-read pre-process file
- Send proof-read, pre-processed file to Ruusbroec/Middle Dutch expert

Home Insert Page Layout Formulas Data Review View Acrobat

Paste Cut Copy Format Painter Clipboard

Arial 10 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting as Table Cell Styles Styles

Insert Delete Format Cells

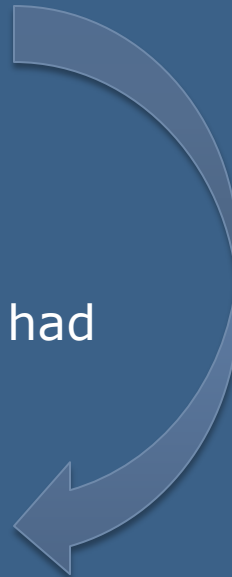
B1 "Siet,

22	A0004	pray aloud, read		bridesmaid, virgin												
23	A0004															
24	A0004															
25	A0005	Dese	brudegon es	Cristus,	ende	menschel	natuere	es	de	bruyt,						
26	A0005		brudegome	Cristus		menschelij	nature			bruu						
27	A0005		n	pn		adj	n			n						
28	A0005		bridegroom			human				bride						
29	A0005															
30	A0005															
31	A0006	die	god	ghemaect	hevet	toe	den	beelde	ende	toe	de	ghelijcke	sijs			
32	A0006		god					beelde				gelikenesse				
33	A0006		n					n				n				
34	A0006							image				likeness, similitude				
35	A0006															
36	A0006															
37	A0007	selfs.	Ende	hi	hadse	gheset	inden	beghinne	in	die	hoochste	stat,	ende			
38	A0007					setten		begin			hooch	stat				
39	A0007					v		n			adj	n				
40	A0007					set		beginning				city, place				
41	A0007															
42	A0007															
43	A0008	in	die	scoonste	ende	in	die	rijkste	ende	in	die	weldichst	van	eertrijcke,		
44	A0008			schone				rike				weeldich		erderike		
45	A0008			adj				adj				adj		n		
46	A0008			fair, lovely				rich				blissful		earth		
47	A0008															
48	A0008															
49	A0009	dat	was	inden	paradise.	Ende	hi	hadde	haer	onderwor	alle	creatueren,				
50	A0009				paradijs					onderwerpen		creature				
51	A0009				n					v		n				
52	A0009				paradise					subject						
53	A0009															
54	A0009															

Fouten

- 16 `nu': geen n., wel adv. (evenzo: 74, 113, 253, 257, 267, 332, 372, 380, 397, 447, 511, 559, 863, 996, 998)
- 17 `binnen': invullen: adv. (evenzo: 144, 276, 304, 559, 587, 688, 695, 725, 729, 737, 826, 889, 978, 1107)
- 20 `goeds': niet `goet, n.' maar wel `goet, adj.'
- 27/8 `geven' staat tweemaal vermeld: 27 en 28
- 52 `meer': het Mnl. W. onderscheidt onbep. telw. (meer, talrijker, een grotere hoeveelheid), bijw. (in hogere mate, intenser) en adj. (groter). Geen gemakkelijk onderscheid. Hierbij mijn opinie voor de hele tekst: **onbep. telw.** (en dus niet op te nemen): 52: meer(1^{ste}), 53 meer(1^{ste}), 385, 463, 536, 537, 563(1^{ste}), 563(3^{de}), 564(2^{de}), 1109(1^{ste}); 1109(2^{de}) **bijw.** (en dus wel op te nemen): 52: meer(2^{de}), 53 meer(2^{de}), 220, 563(2^{de}), 564(1^{ste}), 564(3^{de}), 1108(1^{ste}), 1108(2^{de}) **adj.:** 1149.
- 54 `vele': niet opgenomen in woordenlijst: overeenkomstig het Mnl. W. hier als onbep. .telwoord beschouwd (evenzo: 67, 751)
- 129 `lief ende weerd hebben': overeenkomstig het Mnl. W. worden `lief' en `weert' als aparte adjectieven beschouwd en `hebben' als een apart ww. (`liefhebben' staat alleen als verwijzingslemma in het Mnl. W.). Dus: `lief' adj., `weert' adj., `hebben' v. (zelfde voorstel 764)
- 141a `unst': niet ingevuld: `onste, n.'
- 141b `buten': invullen: adv. (vgl. 17 `binnen', evenzo: 144, 559, 587, 694, 827, 889, 978, 1015)
- 146 `worden': invullen: `werden, v.' (geen hulpww. passief hier)
- 147 `heere': niet ingevuld: `here, n.'
- 150a `sin': op verkeerde plaats
- 150b `formen': geen n. maar v.: lemma `vormen'

Plan C

- Problems:
 - indices follow different norms, none of which is exactly that of MNW
 - Still many mistakes:
 - Separable words
 - Auxiliaries, modals
 - Adj vs nouns e.g. *machteghe*
 - Very, very labor-intensive and slow and the largest text had no index to speak of
 - Initial Hope to “train” a lemmatizer or at least corrector
 - But how to evaluate against different standards?
- 

Plan C

	# tokens	index
Dat rijcke der ghelieven	30 689	+
Die geestelike brulocht	46 607	+
Vanden blinkenden steen	11 901	+
Vanden vier becoringhen	4 307	+
Vanden kerstenen ghelove	5 386	+
Vanden geesteliken tabernakel	128 585	±
Vanden seven sloten	13 666	-
Van seven trappen	14 964	+
Een spiegel der ewigher salicheit	28 608	+
Boecksken der verclaringhe	7 004	-
Vanden XII beghinen	70 220	+
Brieven (7 + 1)	5 677	+
Total - annotated	96 000	

Plan D

Kestemont, Daelemans & De Pauw (2010) *Weigh your words – Memory-Based Lemmatization for Middle Dutch*. Literary and Linguistic Computing.

- Lemmatize and part-of-speech tag Ruusbroec data
- Advantages:
 - Consistent lemmatization (!) and part-of-speech tagging
 - Everything is lemmatized (no missing lemmas)
- Disadvantages:
 - Not an exact match in terms of language
 - Some problems persist: separables, ambiguity, auxiliaries, ...
 - Did I have someone annotate 96k words for nothing?

Demo

<http://aflat.org/rb>

On-going Work

- Construct lemma -> word forms list
- Construct word form -> lemma list

- Correct these lists, rather than actual texts

Future Work

- Evaluate!
 - Calculate number of manual corrections during Plan C
 - Map manually annotated lemmas to MNW for evaluation of Plans A, B and D
- Better interface with admin-privileges for editing remaining errors

Conclusions

- Careful planning?
- How to get gold standard?
- Dead languages: get a real expert in!
- Props to everyone working on NLP for Medieval Dutch