Exploring Language Identification Techniques for Dutch Folktales

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune,
Franciska de Jong, Theo Meder

LREC '12 workshop:
Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects
26 May 2012, Istanbul, Turkey

The Dutch folktale database

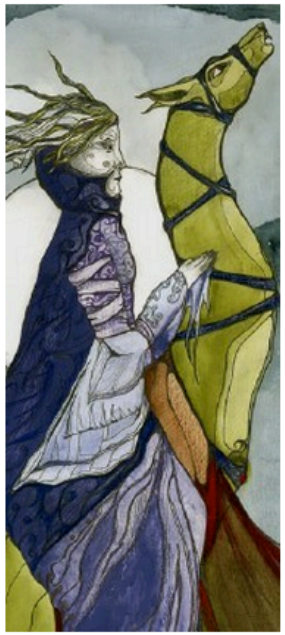
- Maintained by the Meertens Institute since 1994
- > 40,000 Dutch folktales, collected since 19th century
  - Fairy tales
  - Traditional legends
  - Urban legends
  - Jokes, riddles
  - Personal narratives
- In Dutch, Frisian, old Dutch and Dutch dialects

verhalenbank.nl

Home → Databanken

Nederlandse Volksverhalenbank

Toelichting



Welkom!

De Nederlandse Volksverhalenbank bevat momenteel **42122** verhalen.
U kunt hier zoeken naar historische en hedendaagse sprookjes, sagen, legenden, moppen, raadsels en Broodje Aap-verhalen.

zoektermen: zoek toelichting
(vul een trefwoord, naam of regio in, of een combinatie hiervan)
Wilt u geavanceerd zoeken? Wilt u zelf een verhaal aanleveren?

Bekijk het lexicon

Bekijk de recent toegevoegde verhalen

We hebben ons best gedaan om alle regels met betrekking tot privacy en auteursrecht in acht te nemen. Mocht u niettemin van oordeel zijn dat er in strijd is gehandeld met deze regels, neemt u dan contact op met de coördinator.

Verhalen of afbeeldingen die als schokkend zouden kunnen worden ervaren, zijn afgeschermd. Voor inzage in afgeschermde gegevens kunt u slechts op het Meertens Instituut zelf terecht.

Heeft u vragen? Mailt u dan naar de beheerder van de Volksverhalenbank.

Coördinator: Theo Meder
Beheerder: Marianne van Zuijlen
Programmeur/Ontwerp: Maarten van der Peet, Matthijs Brouwer
Illustraties © Minke Priester

Volksverhaal

Cinderella in frisian

Assepoester wie in styfbern. Sy moest it allerswierste wurk dwaen. Har beide susters wienen altyd pronkmoai yn 'e klean, mar Assepoester roan yn fodden om.

Sy hie dêr in min leven. Har styfmem snaude altyd op har om. Yn 'e keuken moest se it fjûr ûnderhâlde en de jiske fuortbringe en sa. Dêrom neamden se har Assepoester.

Op in kear woarde bikend makke, dat de koaning, dy socht in frou. Alle jonge fammen woarden útnoadige om op it paleis to kommen, hwant dêr soe dânse wurde. Trije jounen oanien.

…

(Belgian) jokes

Komt een Belg bij de Opel-dealer, om een proefrit te maken in de nieuwe Astra. Hij is nog geen vijf minuten de deur uit, of hij komt duwend achter de auto terug met een kapotte versnellingsbak. De dealer verzint een of andere smoes ("Foutje van de nieuwe monteur") en geeft de Belg een andere nieuwe Astra mee. Vier minuten later komt hij lopen binnen; de auto staat om de hoek met... een kapotte versnellingsbak. Nu krijgt de dealer toch wel enige argwaan, dus hij besluit mee te rijden bij de volgende proefrit. Dus als ze wegrijden, vraagt hij de Belg wat hij precies deed. De Belg: "Awel, ik rijd weg in z'n 1, trek rustig 't hoekske om en schakel door naar 2. Voorbij het stoplicht zet ik 'm in 3, en op dit stukje rechte weg schakel ik in rap tempo door naar 4, en dan 5. En dan wil ik echt 't uiterste, dus zet ik 'm in den "R" van racen..."

But also urban legends

Een vrouw wordt hinderlijk gevolgd door een auto die met groot licht seint. De vrouw vermoedt dat de bestuurder van de auto haar iets aan wil doen en probeert hem van zich af te schudden. Via de mobiele telefoon stelt de vrouw zich in verbinding met haar vader. De vrouw vertelt dat ze achtervolgd wordt door iemand die waarschijnlijk kwaad in de zin heeft. De vader maant de vrouw naar een benzinestation te rijden, waar hij zelf ook naar toe zal komen. Bij het benzinestation blijkt dat een moordenaar zich al die tijd op de achterbank van de auto van de vrouw schuilgehouden heeft. De bestuurder van de auto die de achtervolging inzette heeft de vrouw middels lichtseinen op de aanwezigheid van de maniak willen attenderen.

Each story is richly annotated

- Manual metadata
  - Story type (e.g. variations of Cinderella have the same type)
  - Summary (Dutch)
  - Keywords (Dutch)
  - Named entities
  - Language
  - …
- A slow and expensive process

FACT:
Folktales as Classifiable Text

- Goal: Develop a user-friendly platform for the annotation and analysis of folktales
  - Develop techniques to speed up metadata annotation (i.e. automatic keyword assignment, language detection, proper name extraction, summarization)
- http://elab-oralculture.nl/FACT

This work: language identification

- How well do standard language identification methods perform on the Dutch Folktale database?

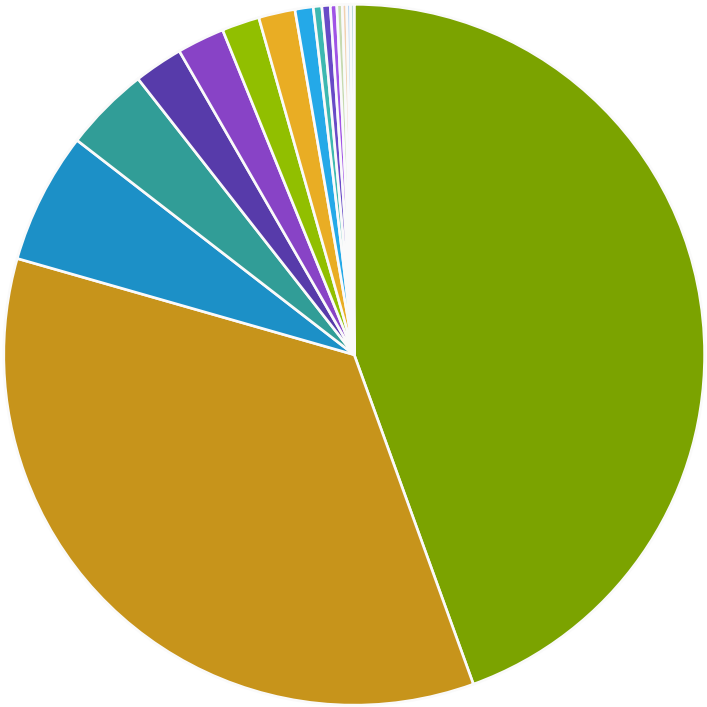
Assepoester wie in styfbern. Sy moest it allerswierste wurk dwaen. Har beide susters wienen altyd pronkmoai yn 'e klean, mar Assepoester roan yn fodden om.
Sy hie dêr in min leven. Har styfmem snaude altyd op har om. Yn 'e keuken moest se it fjûr ûnderhâlde en de jiske fuortbringe en sa. Dêrom neamden se har Assepoester.
Op in kear woarde bikend makke, dat de koaning, dy socht in frou. Alle jonge fammen woarden útnoadige om op it paleis to kommen, hwant dêr soe dânse wurde.
Trije jounen oanien.
…

- Standard Dutch
- 17th century Dutch
- Middle Dutch
- **Frisian**
- Gronings
- ….

Collection statistics (after filtering)

Language	Doc. count
Frisian	17,347
Standard Dutch	13,632
17th century Dutch	2,361
Standard Dutch mixed	1,538
Flemish	882
Gronings[1]	854
Noord-Brabants[1]	677
Middle Dutch	656
Liemers[1]	328
Waterlands[1]	153
Drents[1]	150
Gendts[1]	116
English	97
Overijssels[1]	80
Zeeuws[1]	68
Dordts[1]	64
Total (16 languages)	39,003

[1] Dutch dialects



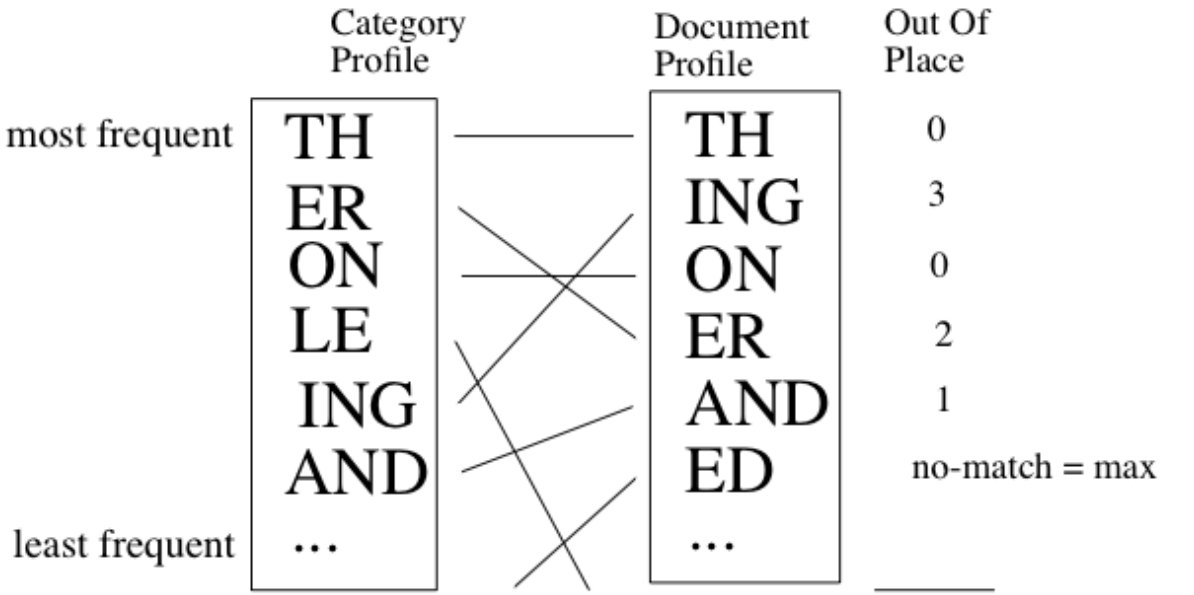
Approach

- Compare supervised learning methods
  - Baseline: TextCat [Cavnar and Trenkle, 1994]

    vs
  - Variations of cosine similarity [inspired by Baldwin and Lui, 2010]
- Stratified 10-fold cross validation
- Evaluation measures: Precision, Recall and F-Measure
  - Macro (per language) and micro (per document) averages

TextCat baseline

- Out of place distance between language profile and text to classify
  - Top 300 character n-grams (n=1..5)



[Cavnar and Trenkle, 1994]

_	2962673
e	957895
n	511681
i	398950
t	387998
a	382483
r	358228
o	320287
s	292793
e_	283662
d	273183
n_	258484
k	185684
t_	180721
_d	170031

Variations on cosine similarity (1)

- Represent text as a vector:

$$\vec{d} = (f_{w_1}, \ldots, f_{w_n})$$

$$d(\text{“an an example”}) = (f_{\text{cows}}, f_{\text{an}}, f_{\text{example}}) = (0, 2, 1)$$

- Determine cosine between language prototype/neighbour and text to classify

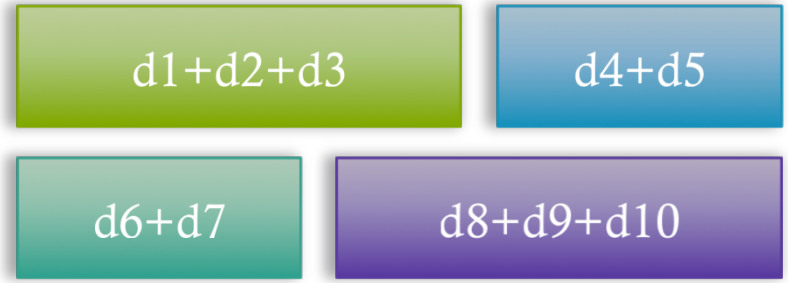
$$\cos(\vec{d_1}, \vec{d_2}) = \frac{d_1 \cdot d_2}{|d_1||d_2|}$$

- Variations:
  - Nearest prototype/neighbour
  - Representation unit: character n-grams (1,2,3,1-4), words
  - Number of features: 100, 500, 1000, all

Variations on cosine similarity (2)

Nearest Prototype

- One large document per language
- Compare to prototype



Nearest Neighbour

- Keep training documents separate
- Compare to each document

Text to classify

Results

Character n-grams	# Features	*Nearest prototype* Precision	Macro Recall	F	Micro P/R/F	*Nearest neighbour* Precision	Macro Recall	F	Micro P/R/F
TextCat		**0.542**	0.676	**0.527**	**0.799**	-	-	-	-
Cosine									
n = 1	all (59)	0.234	0.489	0.243	0.498	0.404	0.419	0.407	0.781
n = 2	100	0.356	0.572	0.365	0.577	0.564	0.525	0.531	0.845
	500	0.405	0.598	0.410	0.597	0.629	0.562	0.579	0.864
	1000	0.406	0.599	0.410	0.598	0.631	0.564	0.581	0.865
	all (1,630)	0.406	0.599	0.410	0.598	0.631	0.564	0.581	0.865
n = 3	100	0.340	0.547	0.338	0.569	0.478	0.494	0.475	0.819
	500	0.451	0.628	0.449	0.630	0.606	0.565	0.561	0.855
	1000	0.484	0.635	0.475	0.630	0.628	0.583	0.582	0.863
	all (17,894)	0.503	0.643	0.490	0.631	0.664	0.598	0.606	0.874
n = 4	100	0.309	0.525	0.323	0.583	0.449	0.408	0.418	0.804
	500	0.375	0.632	0.400	0.637	0.588	0.521	0.540	0.852
	1000	0.376	0.654	0.409	0.641	0.621	0.543	0.568	0.864
	all (112,419)	0.403	**0.693**	0.442	0.656	**0.702**	0.584	0.624	0.886
n = 1...4	100	0.289	0.544	0.309	0.562	0.526	0.516	0.514	0.837
	500	0.354	0.607	0.382	0.638	0.585	0.564	0.567	0.866
	1000	0.372	0.624	0.401	0.658	0.611	0.582	0.588	0.874
	all (132,002)	0.400	0.650	0.431	0.687	0.669	0.601	0.624	**0.887**
words	100	0.369	0.650	0.394	0.643	0.474	0.490	0.475	0.828
	500	0.326	0.560	0.338	0.600	0.612	0.581	0.587	0.862
	1000	0.366	0.638	0.389	0.637	0.627	0.591	0.601	0.867
	all (174,180)	0.373	0.659	0.400	0.649	0.675	**0.609**	**0.630**	0.883

Results Q&A (1)

- *How is the best performance?*
  - Mediocre in comparison to other language id tasks (Baldwin & Lui: 0.729 Macro F with 67 languages)

	Macro F	Micro F
TextCat	0.527	0.799
Cosine, nearest neighbour, all words	**0.675**	**0.883**

Results Q&A (2)

- *How does TextCat compare to Cosine variations?*
  - TextCat is better than Cosine nearest **prototype**
  - Cosine nearest **neighbour** outperforms TextCat

	Macro F	Micro F
Cosine, nearest prototype, 1..4-grams	0.431	0.687
TextCat	0.527	0.799
Cosine, nearest neighbour, all words	**0.630**	**0.883**

Results Q&A (3)

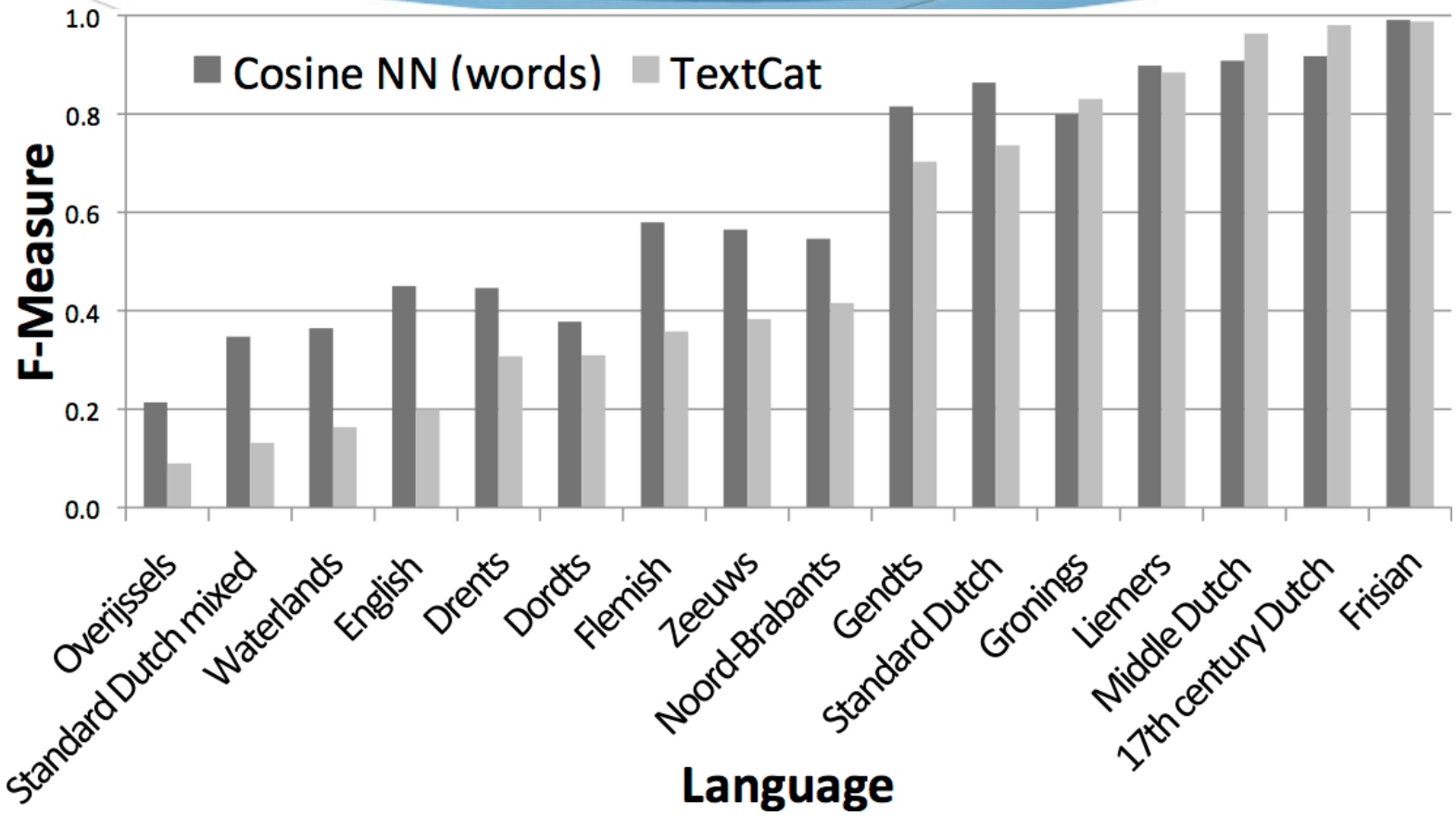
- *What Cosine variation to use?*
  - Use nearest neighbor rather than nearest prototype
  - Use words rather than character n-grams
  - Use as many features as possible

	Macro F	Micro F
Cosine, nn, 3-gram, 100 features	0.475	0.819
Cosine, nn, 3-gram, all features	0.581	0.865
Cosine, nn, words, 100 features	0.475	0.828
Cosine, nn, words, all features	**0.630**	**0.883**

Results Q&A (4)

- *Which languages are easy/hard?*
  - Frisian and old Dutch are easy
  - English performs poor
  - Dialects vary strongly
  - Mixed Dutch performs poor

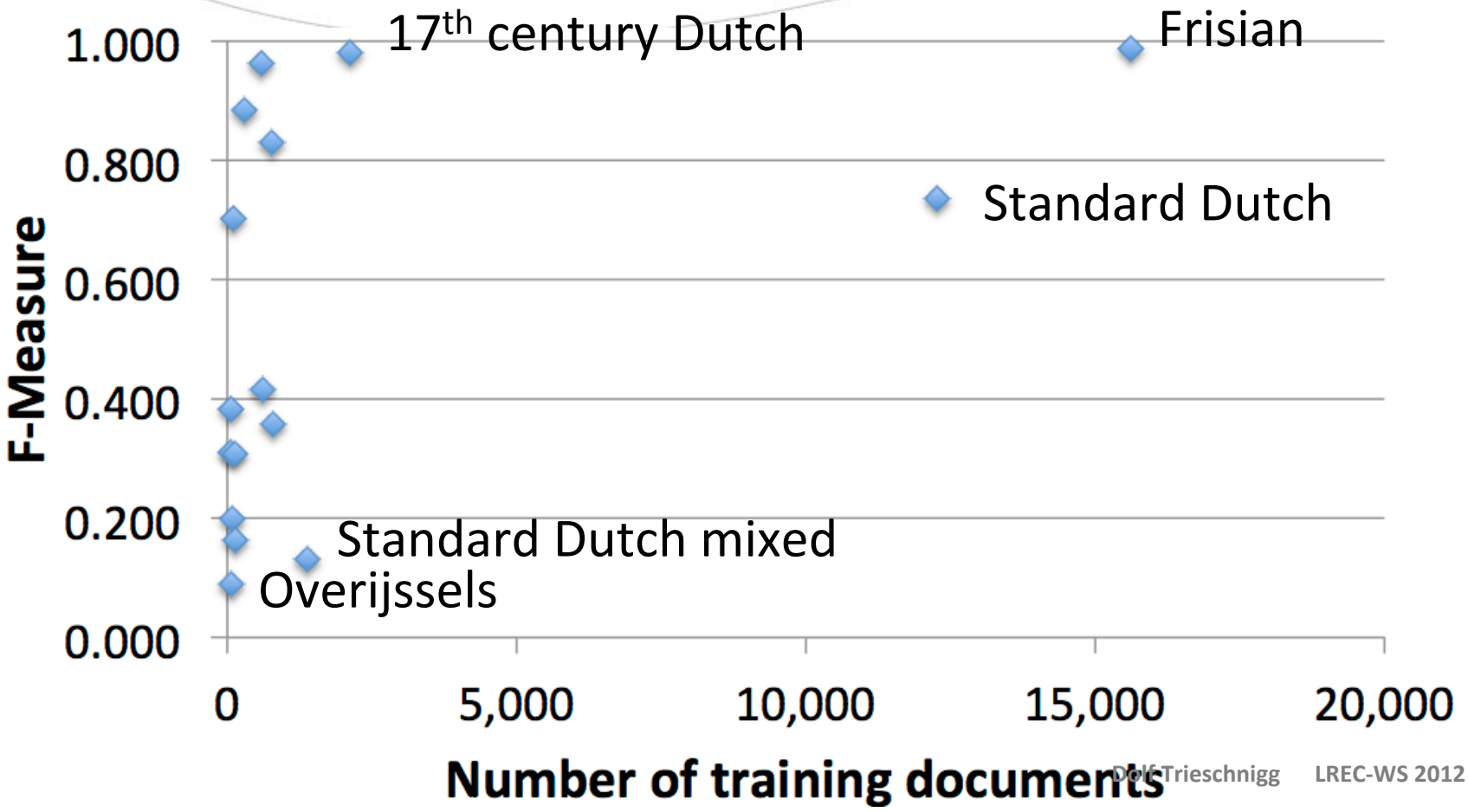
Performance per language



Results Q&A (5)

- *Is performance correlated to amount of training data?*
  - no

Amount of training data vs. performance



Conclusions

- The Dutch folktale database: a large and challenging collection for language identification
- Using standard methods
  - Identification of Frisian and old Dutch documents is easy
  - Identification of Dutch dialects varies strongly
  - Identification of mixed Dutch is hard
- A lot of training data is no guarantee for good results
- Best performance with nearest neighbour cosine similarity based on words

Future work

- Support for multilingual documents (per sentence)
- Assign certainty to classification
- Present “proof”
- Include other helpful non-text features (other metadata fields)
- Use dialect lexicons

Questions?