Het Parsen van Tweets

Kroegcollege ASCI

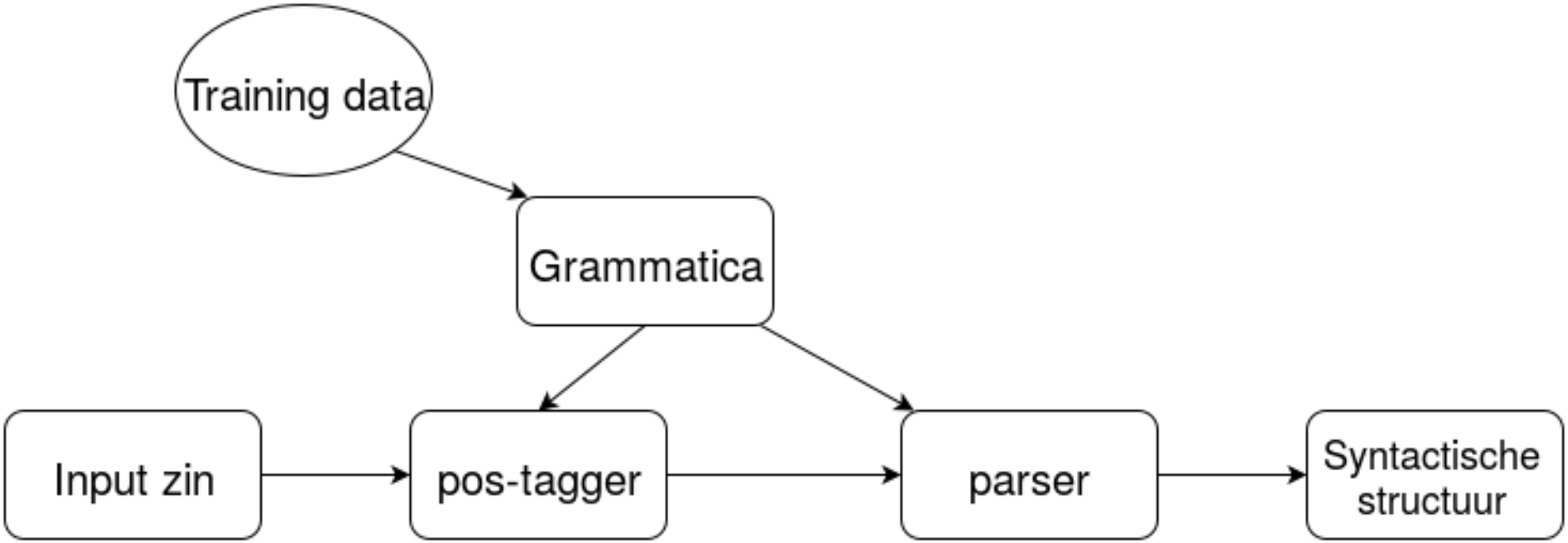
Rob van der Goot

15-10-2015

Table of Contents

Table of Contents

Parsen:


Training data
Grammatica
Input zin
pos-tagger
parser
Syntactische structuur

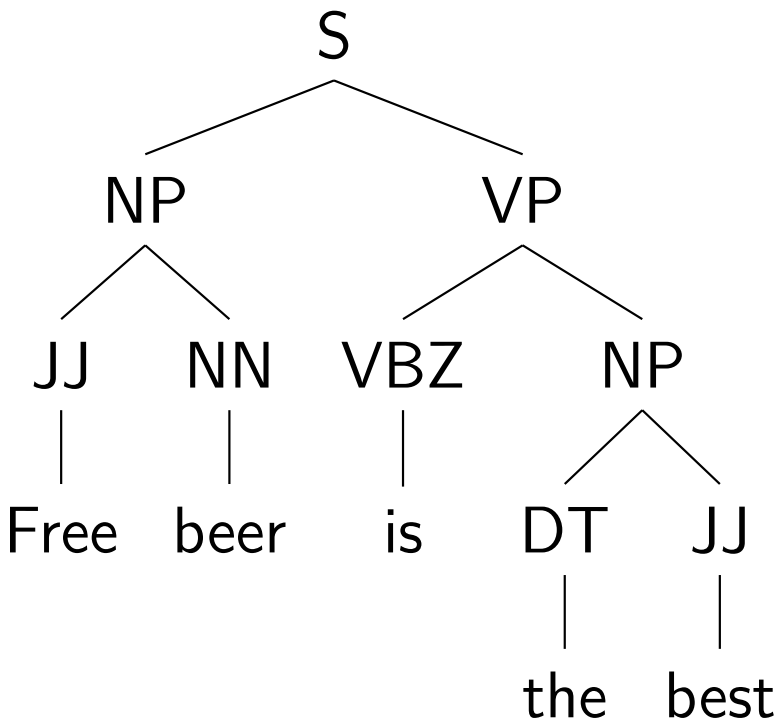
Parsen: Part-Of-Speech Tagging

- Toewijzen van label aan woord (ontleden)
- Op basis van vorige voorkomens van woord
- Op basis van omliggende woorden

Free	beer	is	the	best	.
JJ	NN	VBZ	DT	JJ	.

Parsen: Parsen

- Structuren van nieuwe zinnen (pos sequences) afleiden
- Op basis van structuren van eerder bekeken zinnen



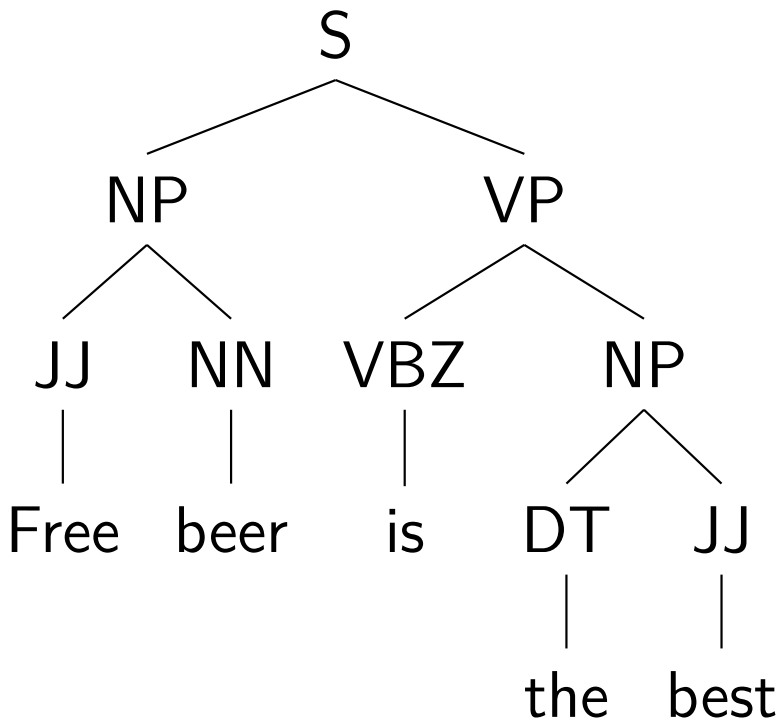
Parsen: Data

Twitter als dataset

- Voordeel: hoeveelheid
- Voordeel: lengte
- Nadeel: lengte
- Nadeel: hoeveelheid geannoteerd
- Nadeel: ruis
- Nadeel: bruikbaarheid

Parsen: Data

Twitter als dataset



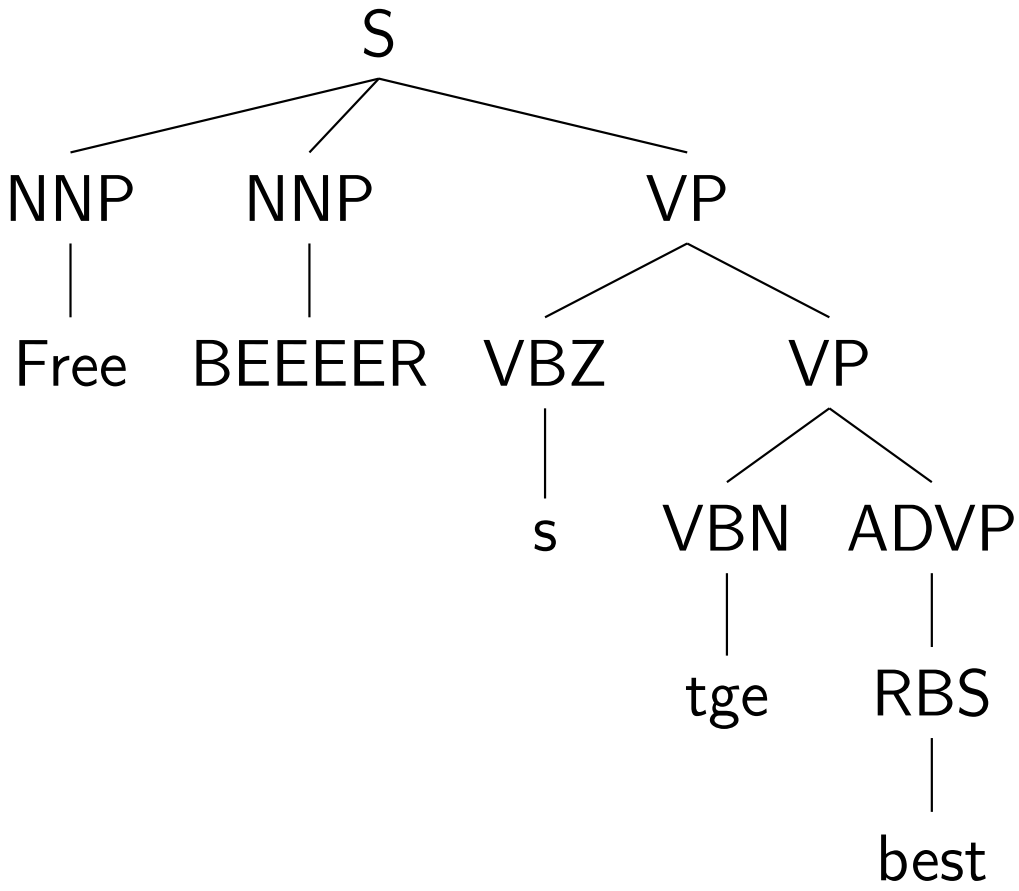


Table of Contents

Previous Work

- Re-training

Previous Work

- Re-training
- Normalizatie

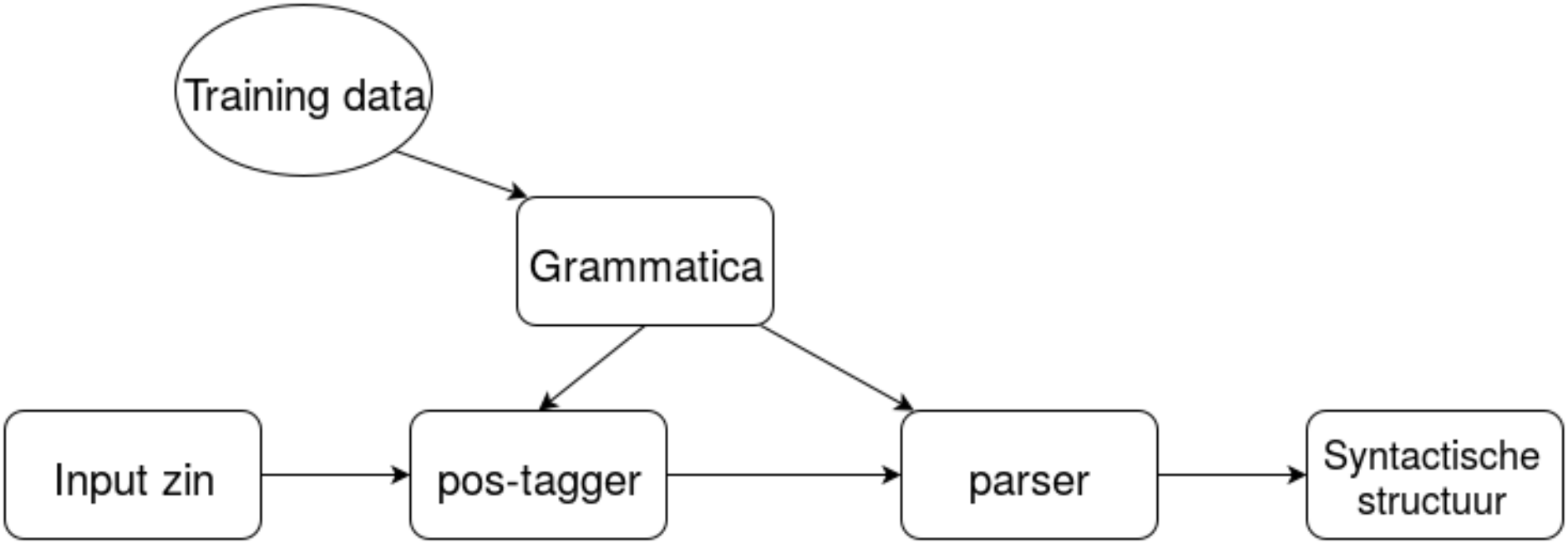
Previous Work

- Re-training
- Normalizatie
- Re-training

Previous Work

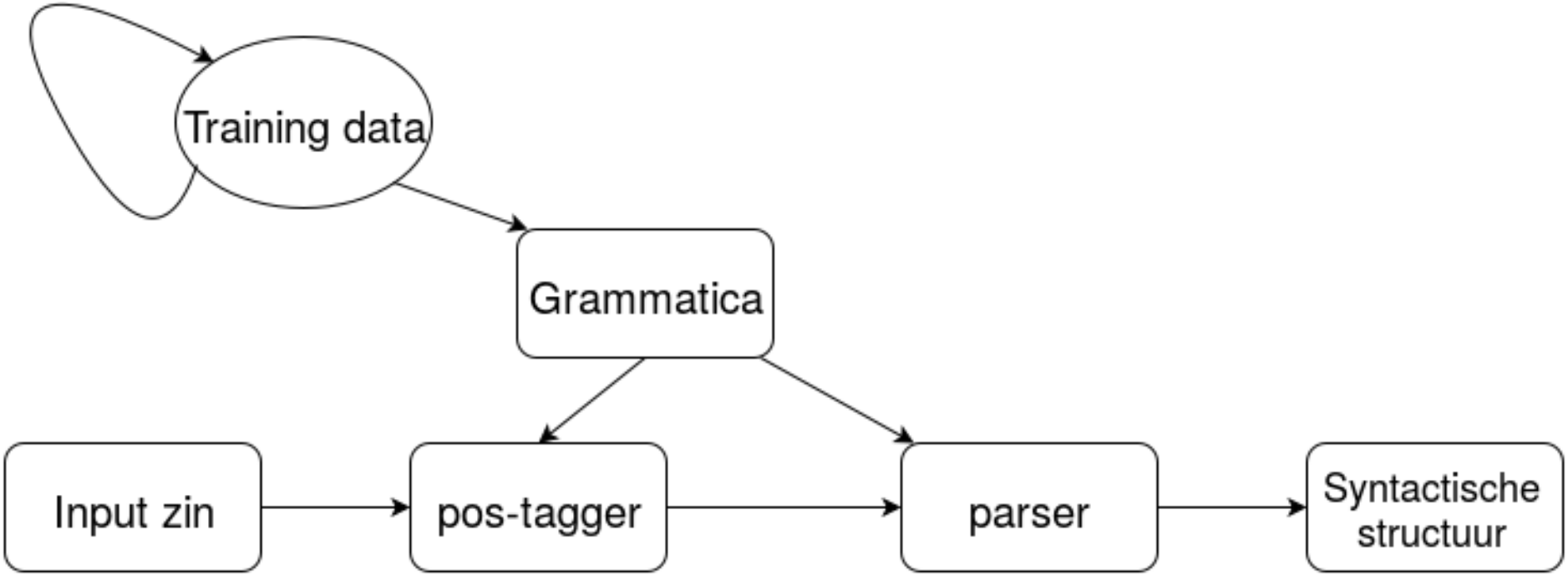
- Re-training
- Normalizatie
- Re-training
- Normalizatie

Previous Work



Previous Work

Re-training



Previous Work

Normalizatie

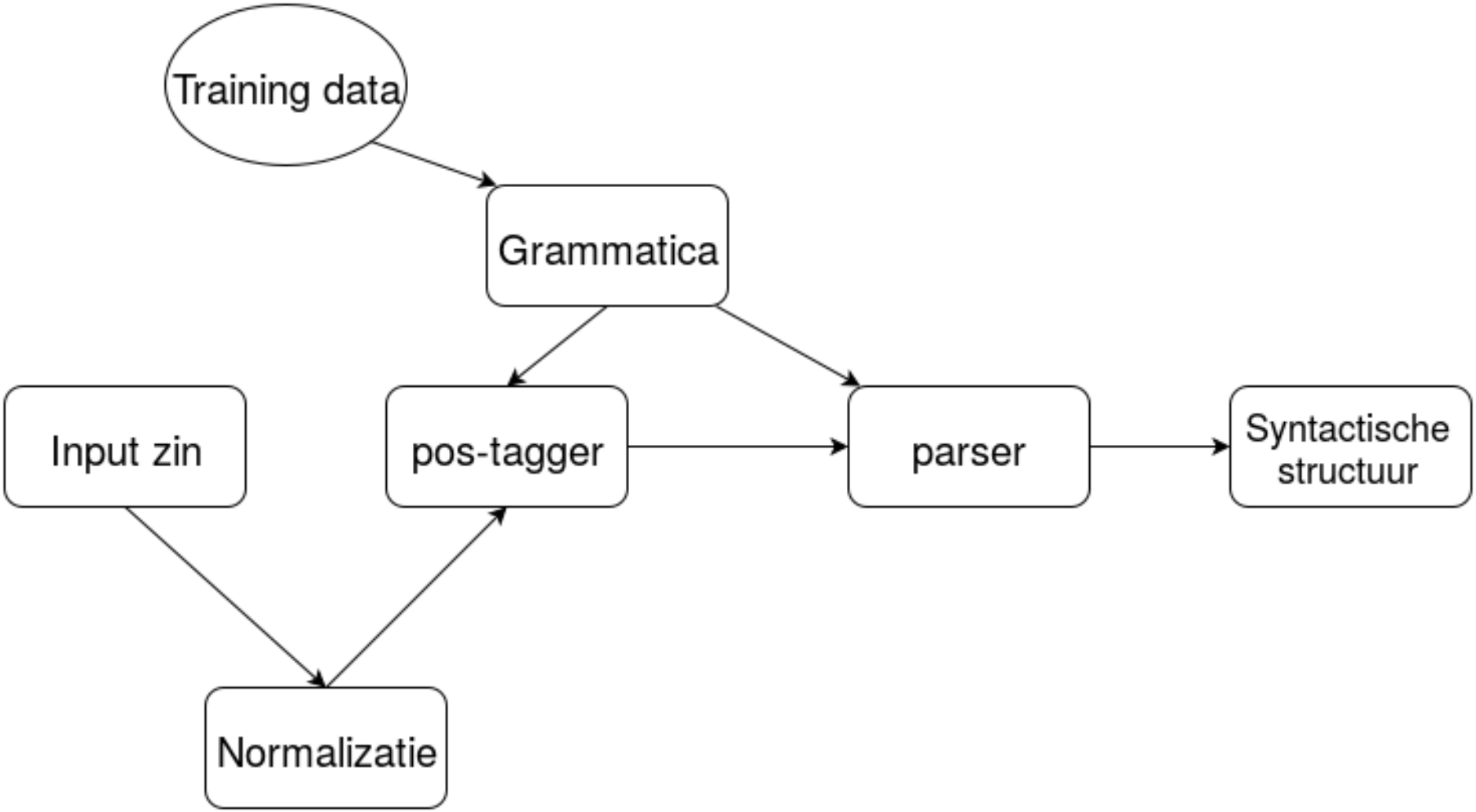
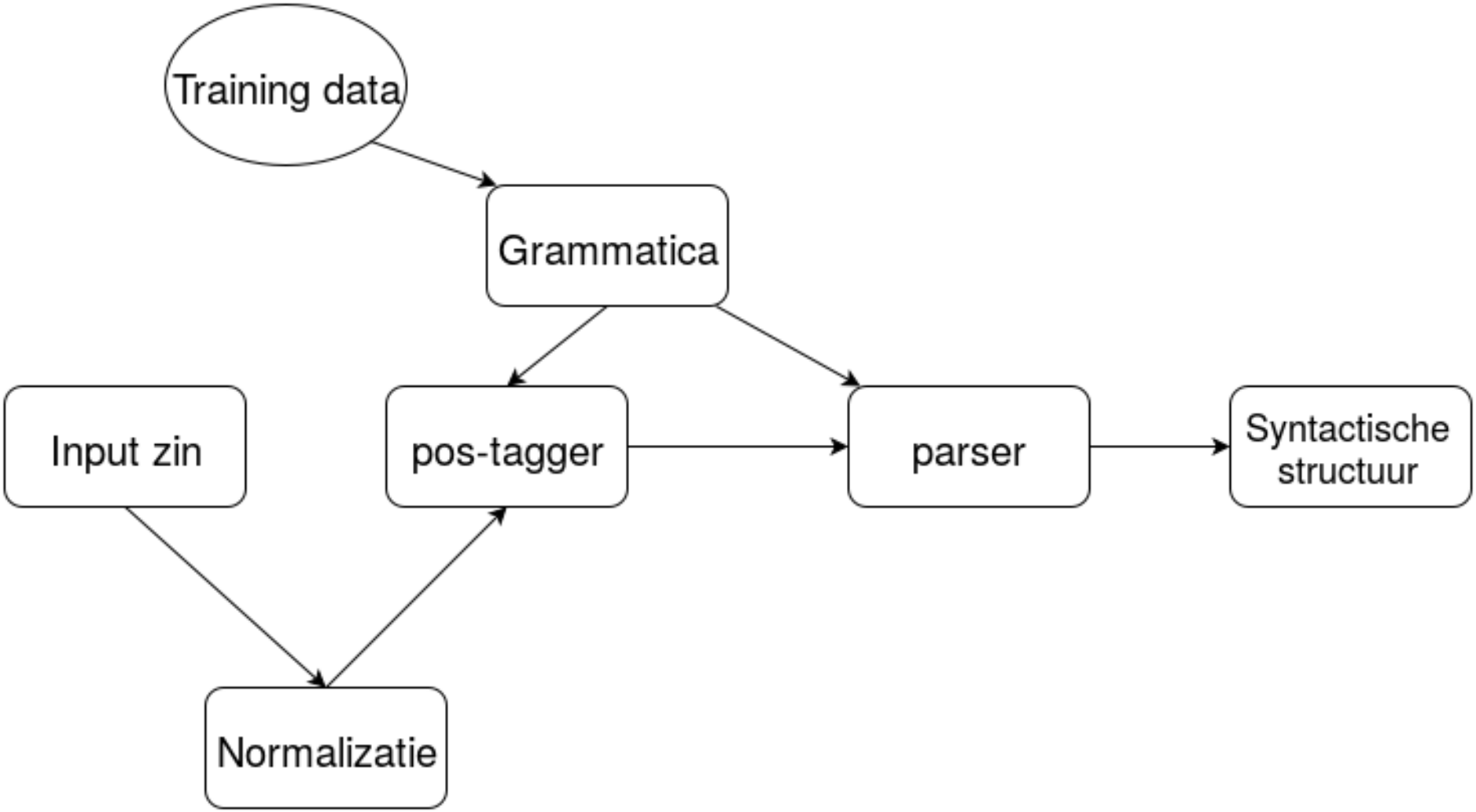


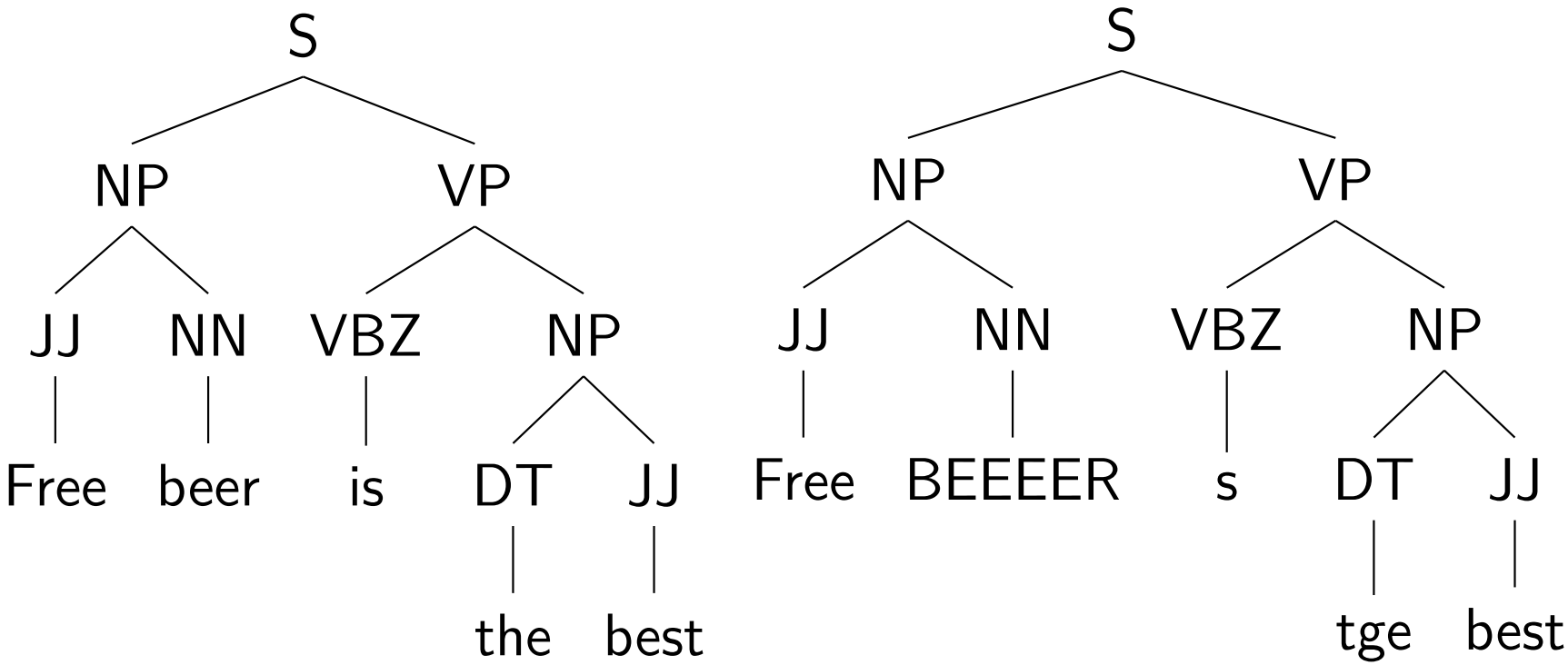
Table of Contents

Normalizatie: Hoe te gebruiken?



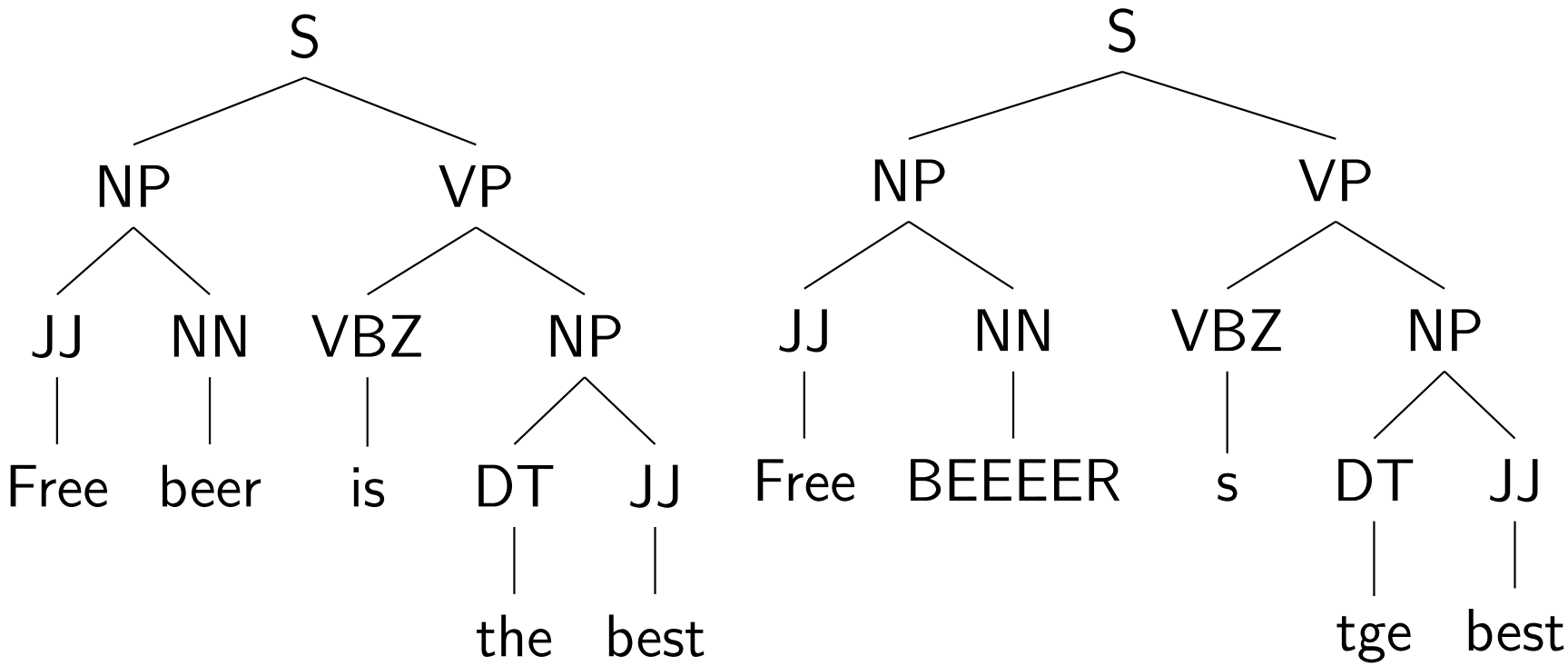
Normalizatie: Hoe te gebruiken?

- Input: "Free BEEEER s tge best" $\mapsto$
- Normalizatie: "Free beer is the best" $\mapsto$
- Pos-tagger: "Free(JJ) beer(NN) is(VBZ) the(DT) best(JJ)" $\mapsto$
- Parser:

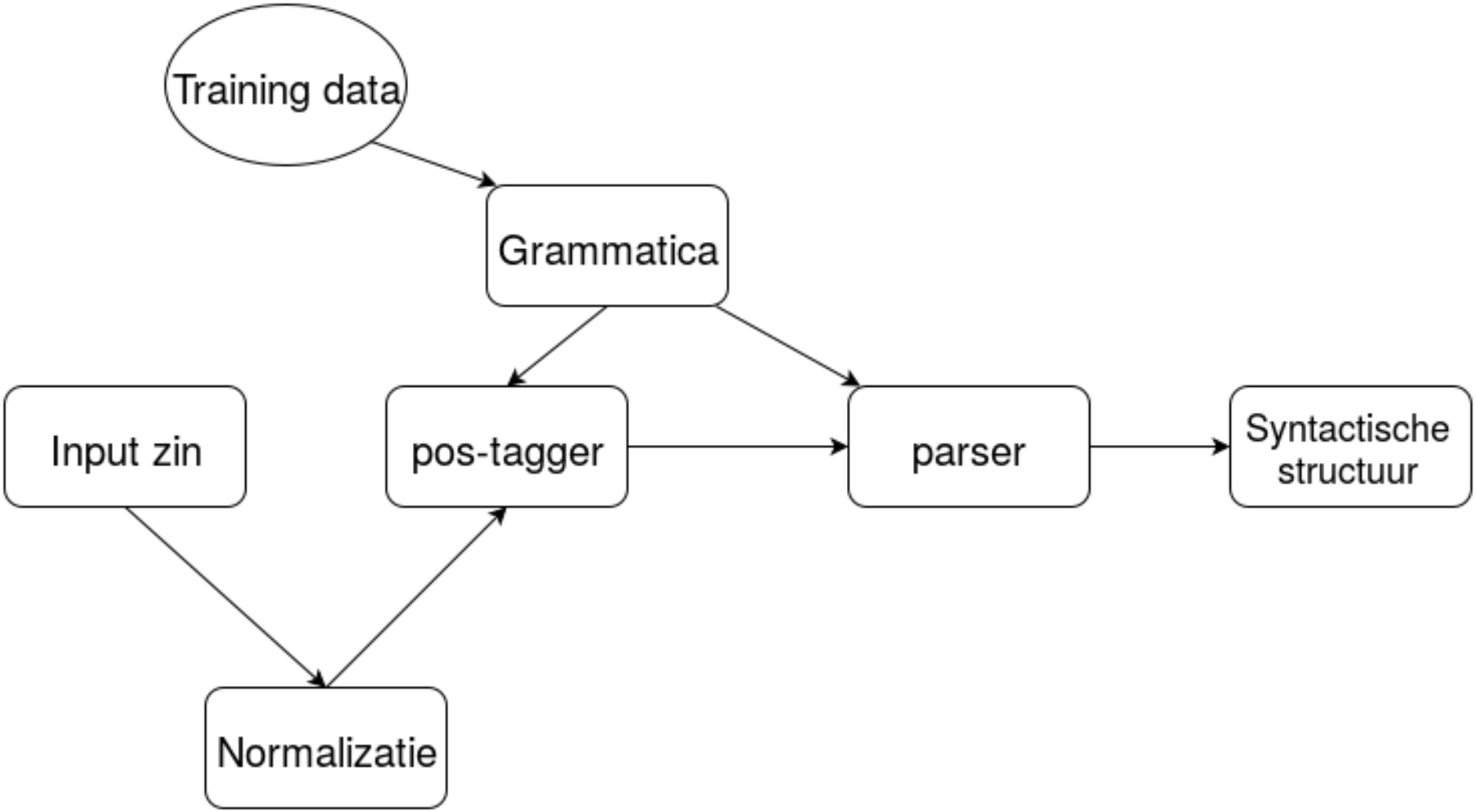


Normalizatie: Hoe te gebruiken?

- Input: "Free BEEEER s tge best" $\mapsto$
- Normalizatie: "Free beer is the best" **(hopelijk!)** $\mapsto$
- Pos-tagger: "Free(JJ) beer(NN) is(VBZ) the(DT) best(JJ)" $\mapsto$
- Parser:



Normalizatie: Hoe te gebruiken?



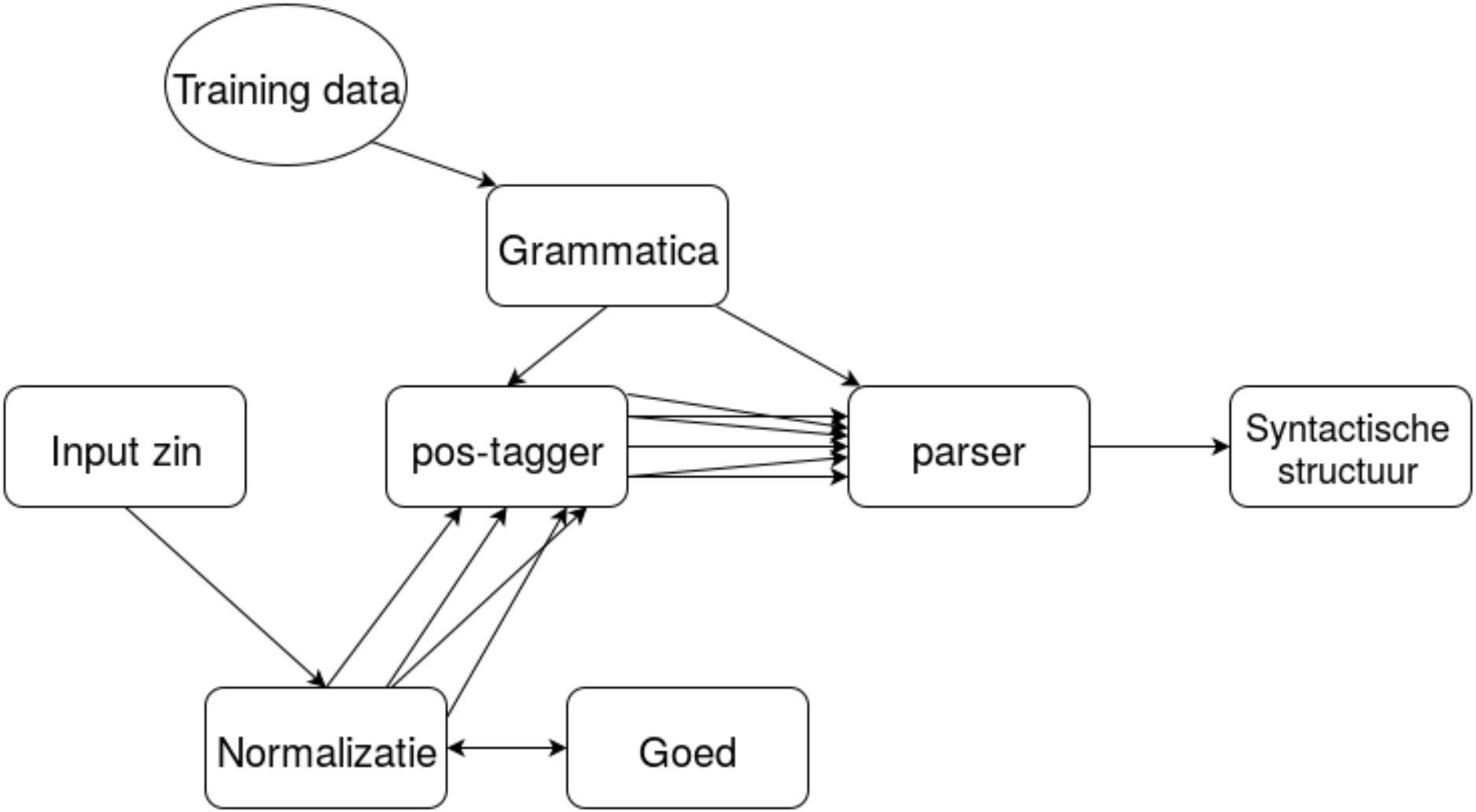
Normalizatie: Hoe te gebruiken?

- Gebruik niet 1 maar meerdere mogelijke normalizaties
- De parser beslist welke het beste past

Normalizatie: Hoe te gebruiken?

- Gebruik niet 1 maar meerdere mogelijke normalizaties
- De parser beslist welke het beste past
- Features van de normalizatie kunnen gebruikt worden in de parser
- Features van de parser kunnen gebruikt worden voor de normalizatie

Normalizatie: Hoe te gebruiken?



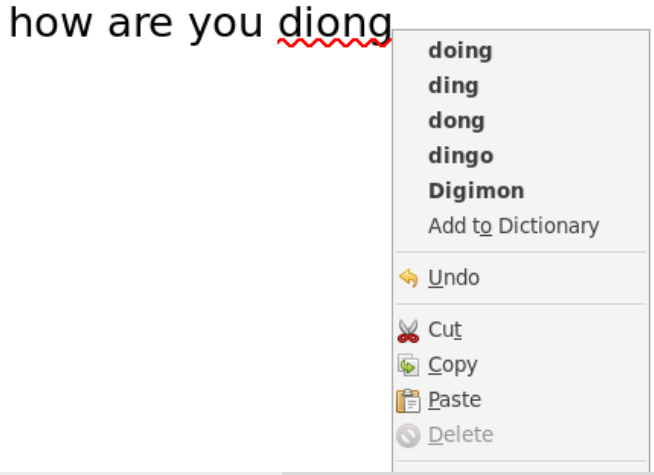
Normalizatie: Methode

- Error Detectie

how are you diong

Normalizatie: Methode

- Error Detectie
- Generatie
- Ranking



Normalizatie: Error Detectie

Oude Methode

- Dictionary lookup

Normalizatie: Error Detectie

Dictionary lookup

Name	Size	Errors	Recall	Precision
Aspell	118,819	0.352	0.795	0.238
Scowl	165,458	0.347	0.779	0.237
Hunspell	48,744	0.880	0.976	0.117
OpenOffice	51,043	0.826	0.917	0.117

- Recall: percentage van de errors gevonden
- Precision: percentage van de gevonden errors, dat ook echt een error is

Normalizatie: Error Detectie

Dictionary lookup

Name	Size	Errors	Recall	Precision
Aspell	118,819	0.352	0.795	0.238
Scowl	165,458	0.347	0.779	0.237
Hunspell	48,744	0.880	0.976	0.117
OpenOffice	51,043	0.826	0.917	0.117

- Recall: percentage van de errors gevonden
- Precision: percentage van de gevonden errors, dat ook echt een error is
- We willen een zo hoog mogelijke recall

Normalizatie: Error Detectie

Dictionary lookup

Normalizatie: Error Detectie

Dictionary lookup

IV/OOV

I am tiret

tired

tire

I am tiret

is amy tired

im aim tire

Normalizatie: Error Detectie

Dictionary lookup

IV/OOV

I am tire

I am tire

is amy tired

im aim tire

Normalizatie: Error Detectie

Dictionary lookup

Hoe kunnen we de recall verhogen?

Name	Size	Errors	Recall	Precision
Aspell	118,819	0.352	0.795	0.238
Scowl	165,458	0.347	0.779	0.237
Hunspell	48,744	0.880	0.976	0.117
OpenOffice	51,043	0.826	0.917	0.117

- Recall: percentage van de errors gevonden
- Precision: percentage van de gevonden errors, dat ook echt een error is

Normalizatie: Error Detectie

Dictionary lookup

Name	Size	Errors	Recall	Precision
Aspell	118,819	0.352	0.795	0.238
Scowl	165,458	0.347	0.779	0.237
Hunspell	48,744	0.880	0.976	0.117
OpenOffice	51,043	0.826	0.917	0.117
Greedy		1.0	1.0	0.115

Normalizatie: Error Detectie

Dictionary lookup

Normalizatie: Generatie

- Lexicale afstand (typos)
- Dingen die hierbuiten vallen.. (afstand in betekenis)

Normalizatie: Generatie

Unintended noise

alchohol	alcohol
alchoholic	alcoholic
daly	daily
dinasour	dinosaur
teh	the

taken from: `http://aspell.net/test/cur/batch0.tab`

Normalizatie: Generatie

Lexicale afstand

- Levehnstein Distance
- Double Metaphone (soundex)

Normalizatie: Generatie

Lexicale afstand

Typo	Correction	Levehnstein	Double Metaphone	Dist.
alchoholic	alcoholic	1	ALKH	0
daly	daily	1	TL	0
dinasour	dinosaur	2	TNSR	0
teh	the	2	T	0

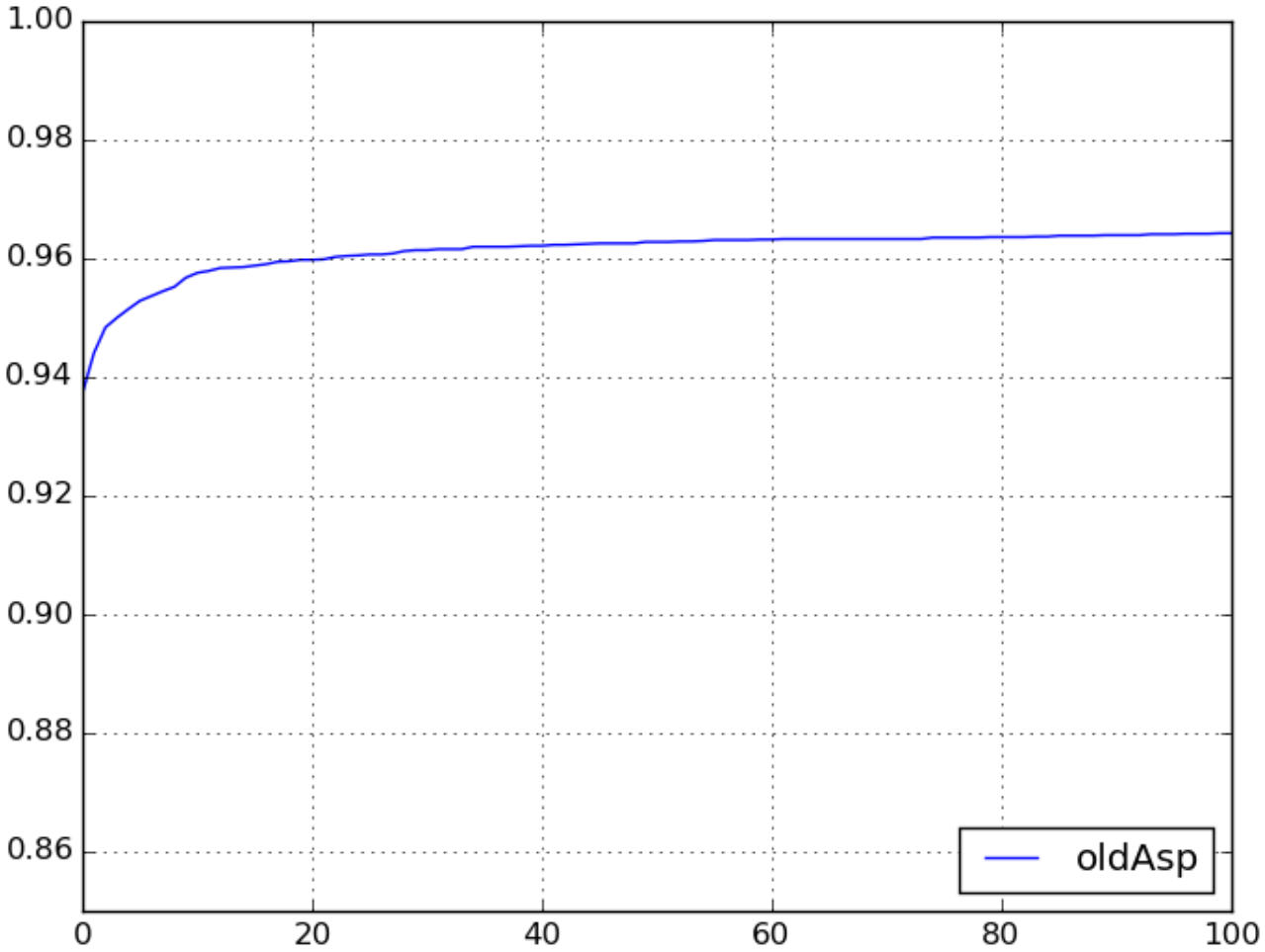
Normalizatie: Generatie

Lexical Normalization = zo goed als opgelost probleem

Normalizatie: Generatie

Lexical Normalization = zo goed als opgelost probleem
Dus gebruik ik aspell

Normalizatie: Generatie



Normalizatie: Generatie

Niet gevonden

```
Every1 everyone
anythinggggg anything
2morrow tomorrow
sum1 someone
granny grandmother
cuz because
apps applications
```

Normalizatie: Generatie

Maar: Lexicale afstanden vinden vooral correcties voor Unintended noise (typos)

Normalizatie: Generatie

Maar: Lexicale afstanden vinden vooral correcties voor Unintended noise (typos)
Dus: Slimmere methode nodig (aanvulling)

Normalizatie: Generatie

Ideeen:

Normalizatie: Generatie

Ideeen:

- Aanpassen van lexicale afstand

Normalizatie: Generatie

Ideeen:

- Aanpassen van lexicale afstand
- Word Embeddings

Normalizatie: Generatie

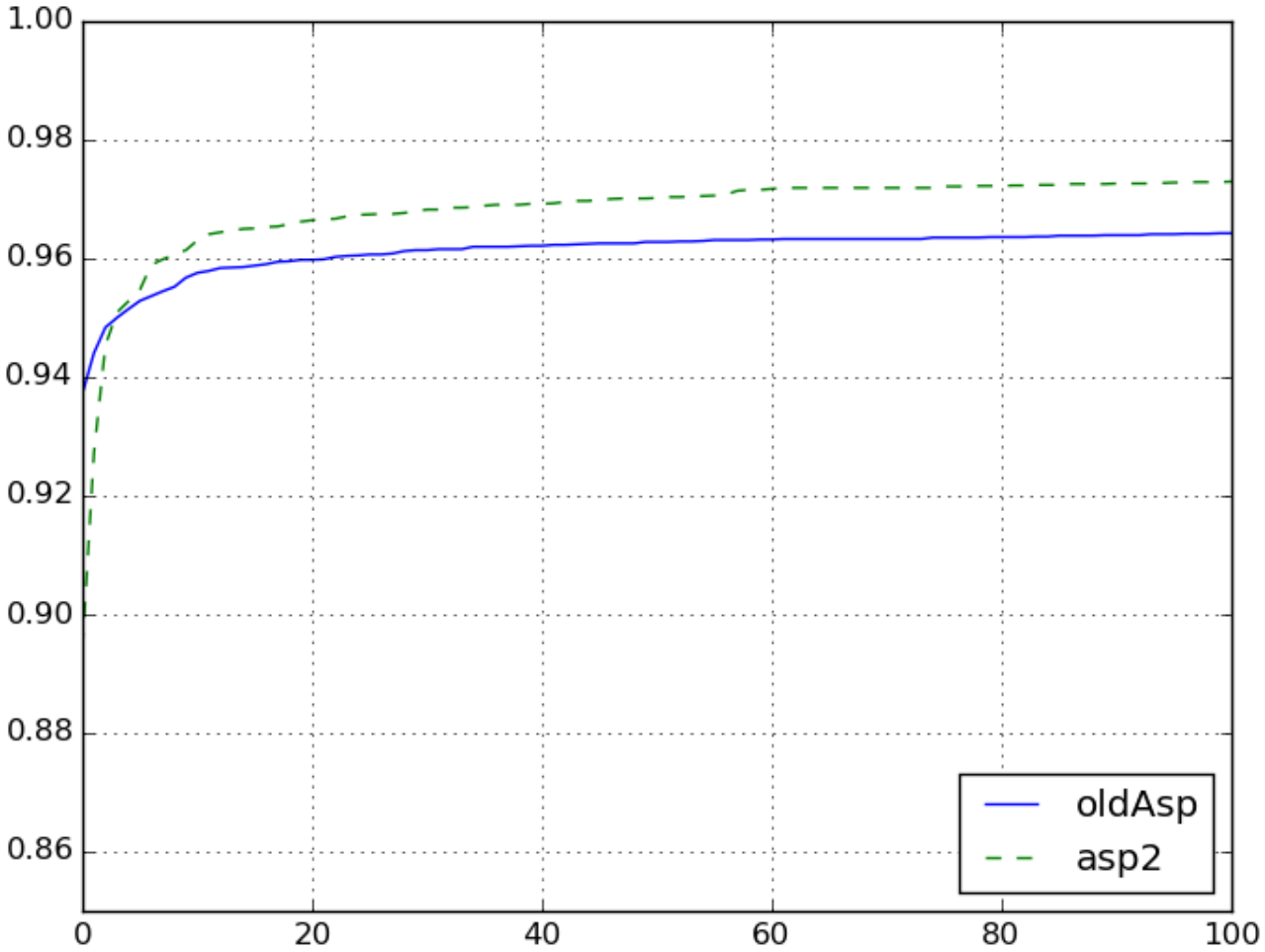
Aanpassen van lexicale afstand

- Verwijderen van meer dan 3 dezelfde opeenvolgende characters
- Gebruik van characters als uitspraak:

$$\begin{array}{lll} 0 & \mapsto & \text{zero, o} \\ 1 & \mapsto & \text{one, i, l} \\ b & \mapsto & \text{bee} \\ c & \mapsto & \text{cee} \end{array}$$

b4 $\mapsto$ beefour $\mapsto$ aspell $\mapsto$ before

Normalizatie: Generatie



A too short introduction to word embeddings

- Representeer woorden als vectors
- Waardes in vector gebasseerd op co-occurences

A too short introduction to word embeddings

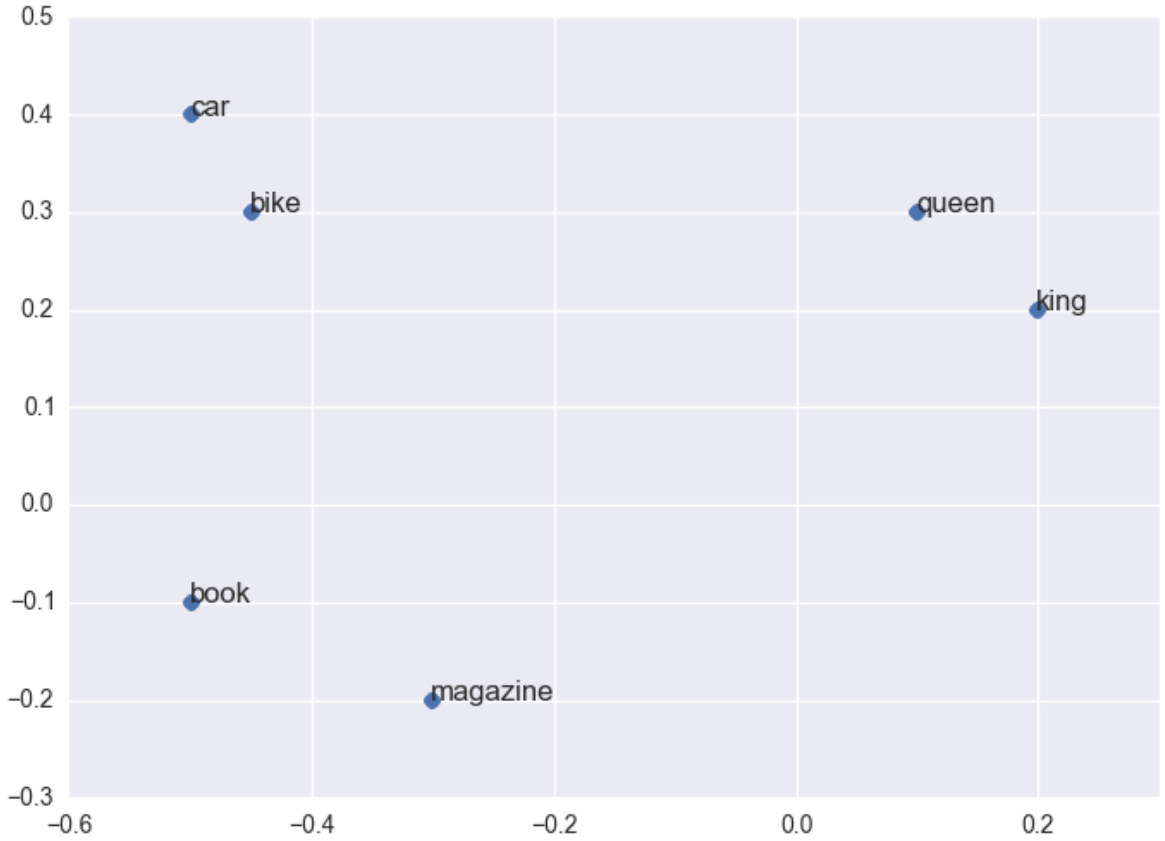
- Representeer woorden als vectors
- Waardes in vector gebasseerd op co-occurences
- Representeer woorden op basis van omliggende woorden

A too short introduction to word-embeddings

because -0.016895 -0.23224 0.13161 -0.060851 -0.27829 -0.38843
1.9953 -0.74172 -0.96761 0.42466 -0.11685 0.44388 -5.3205
-0.20027 0.28929 0.60342 -0.015802 -0.52457 0.64583 -0.76092
0.19877 0.46181 -0.1888 -0.025484 -0.80516
cuz -0.026389 1.3791 0.57634 -0.24492 -0.10887 0.12729 0.72276
-0.54239 -0.62418 0.62131 -1.2408 0.63406 -4.7199 -0.38907
-0.079412 0.61559 0.50521 -0.86192 -0.018521 0.25637 1.0394
0.65304 0.85905 0.48483 0.10626

taken from: `http://nlp.stanford.edu/projects/glove/`

A too short introduction to word embeddings



Taken from: `http://www.folgertkarsdorp.nl/word2vec-an-introduction/`

Normalizatie: Generatie

Waarom?

- Unsupervised: dus geen geanotteerde data nodig
- Effectief
- Wijd inzetbaar

2013	15
2014	70
2015	350

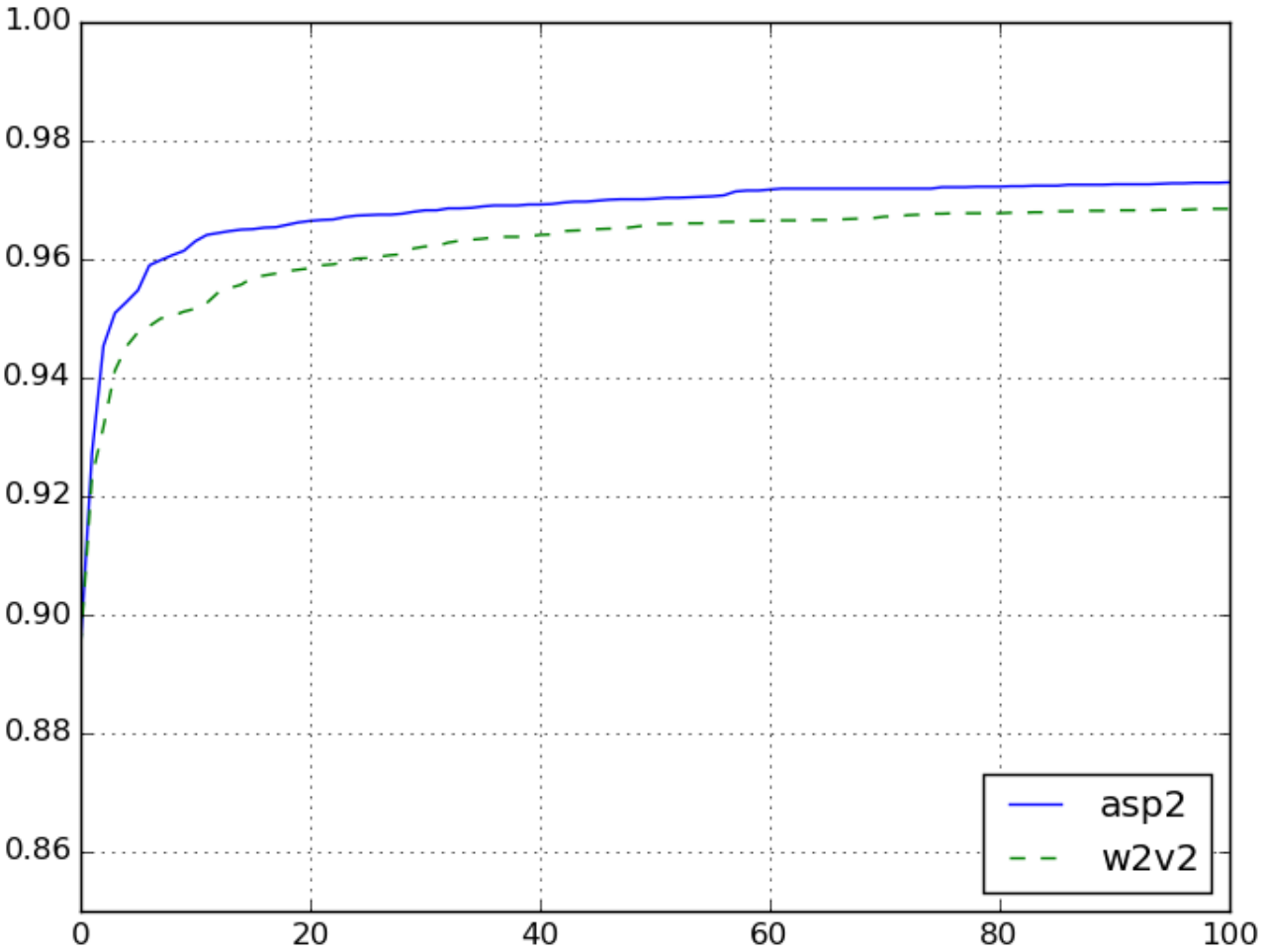
Aantal voorkomens 'word embeddings' in ACL proceedings

Normalizatie: Generatie

Demo

Normalizatie: Generatie

Word embeddings



Normalizatie: Generatie

Newly found:

@	at
pics	pictures
till	until
cuz	because
da	the
granny	grandmother

Normalizatie: Generatie

Hoe verder?

0.97% gevonden, 121 verschillende errors over:

```
night goodnight
cuz cousin
const construction
Schd scheduled
poopsicles popsicles
H8 hate
facebookin facebooking
fina finally
```

Normalizatie: Generatie

Hoe verder?

- Woord.*

Normalizatie: Generatie

Hoe verder?

- Woord.*
- Improve word embeddings

Normalizatie: Generatie

Hoe verder?

- Woord.*
- Improve word embeddings
- 2 step approach

Normalizatie: Generatie

Hoe verder?

- Woord.*
- Improve word embeddings
- 2 step approach
- Enlarge aspell dictionary

Normalizatie: Ranking

Unsupervised

Combineren twee methodes:

- Aspell
- Word embeddings

Normalizatie: Ranking

Unsupervised

Combineren twee methodes:

- Aspell
- Word embeddings
- Ngrams

Normalizatie: Ranking

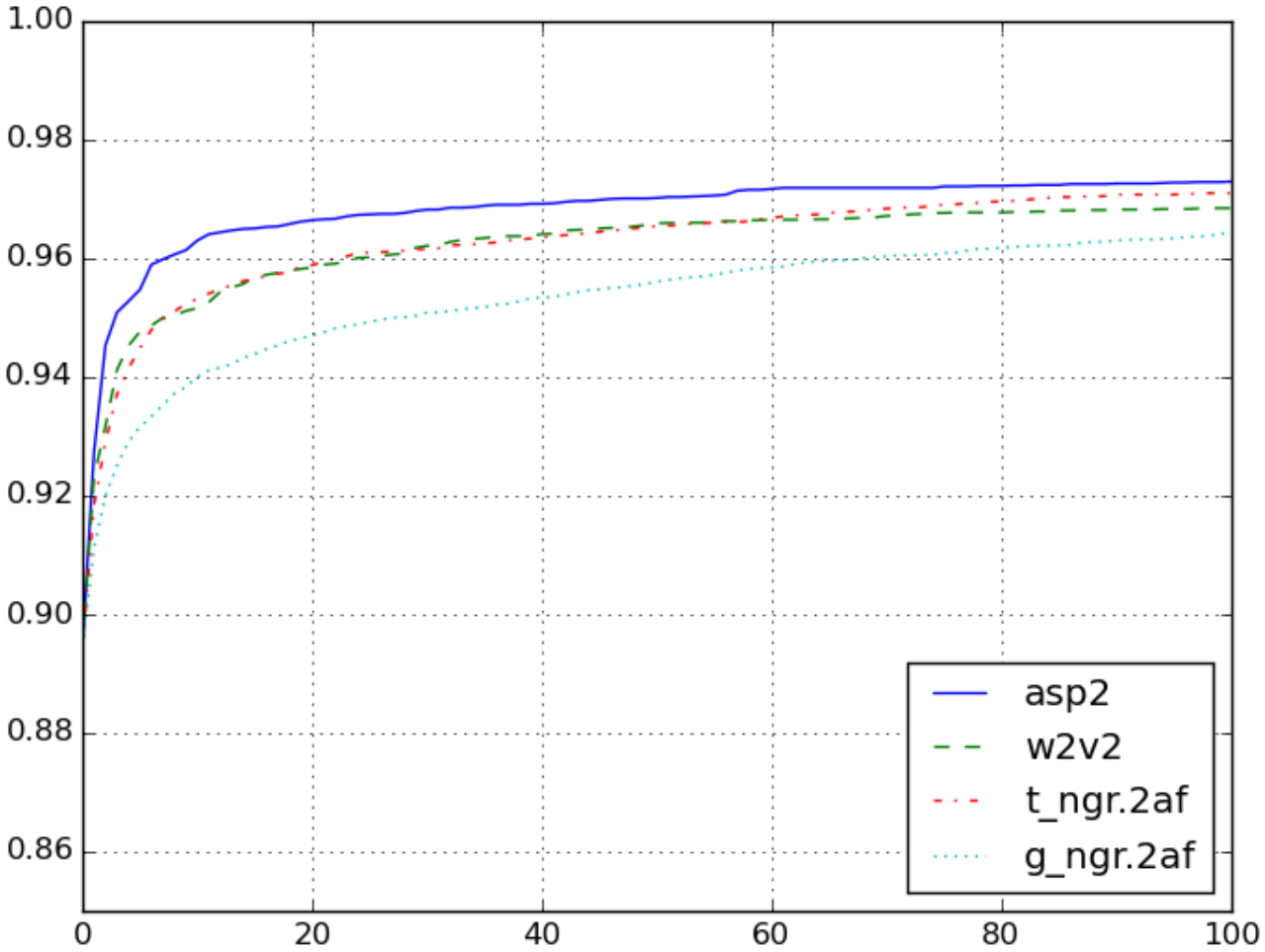
Wat zijn ngrams?

Normalizatie: Ranking

Wat zijn ngrams?

Demo

Normalizatie: Ranking



Normalizatie: Ranking

Normalizatie: Ranking

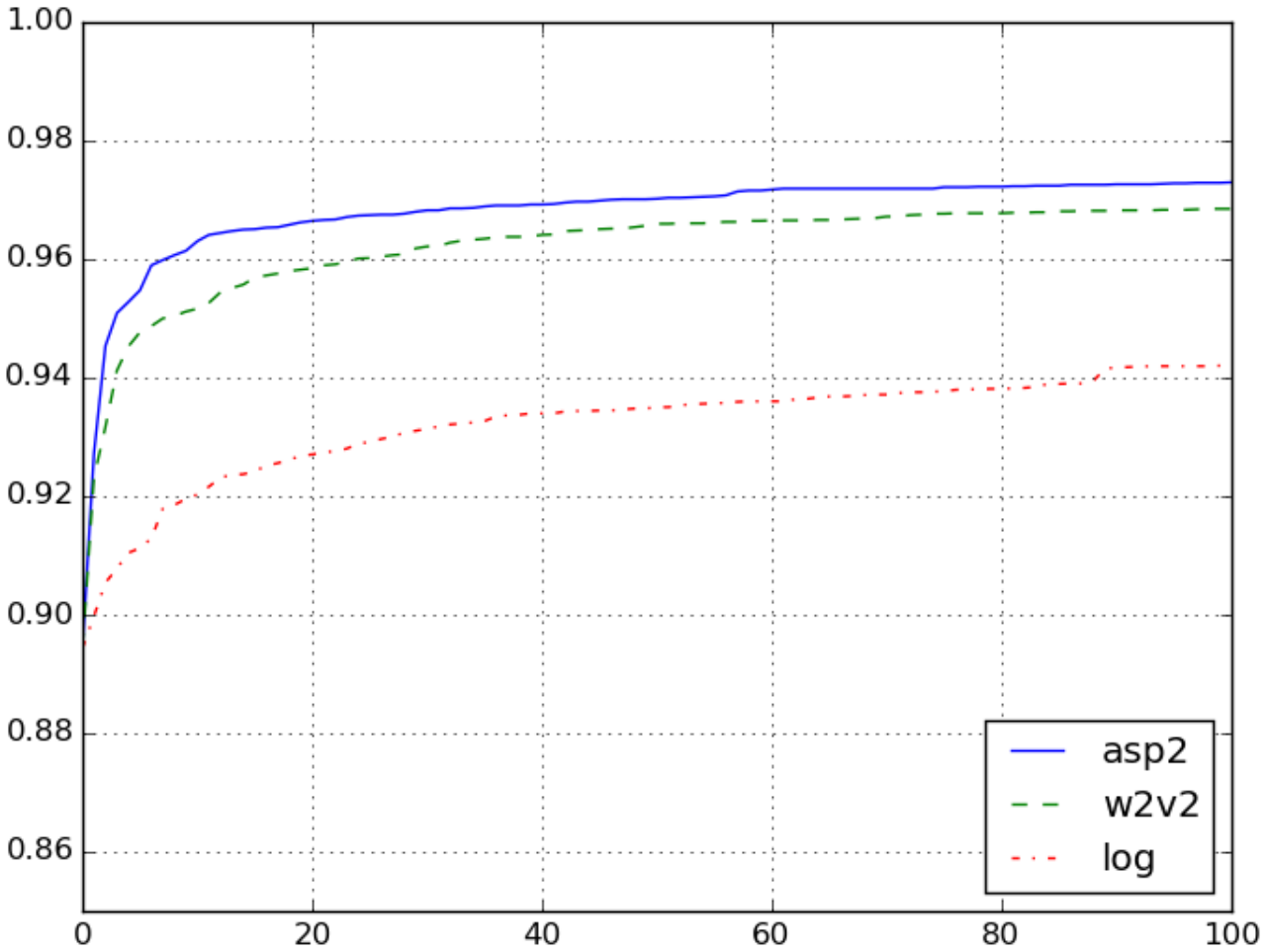
- Verschillende scores samenvoegen tot 1 beslissing:

Normalizatie: Ranking

- Verschillende scores samenvoegen tot 1 beslissing:
- Machine Learning

Normalizatie: Ranking

Eerste poging: Maximum Entropy

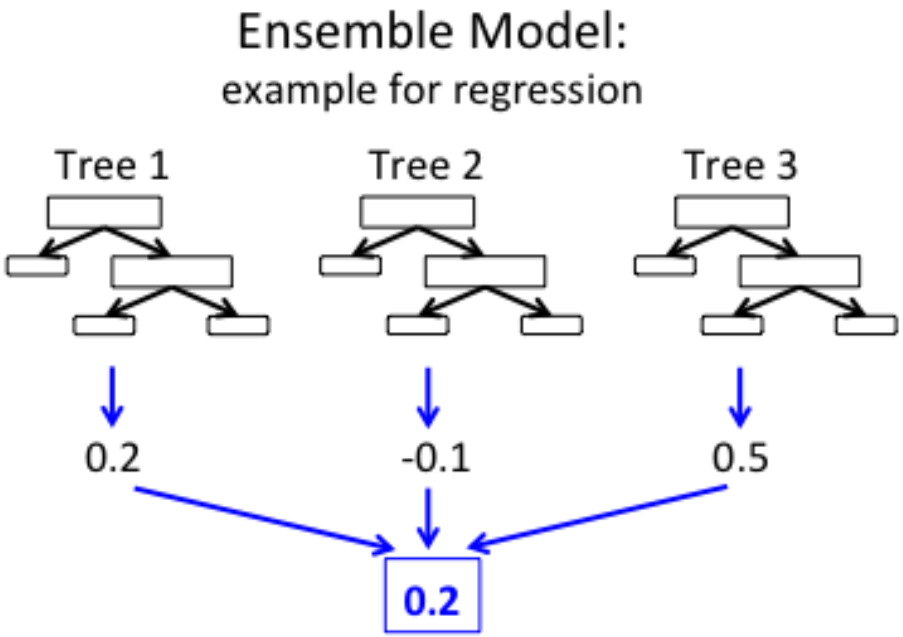


Normalizatie: Ranking

Eerste poging: Maximum Entropy

Normalizatie: Ranking

Tweede poging: Random Forest Regressor

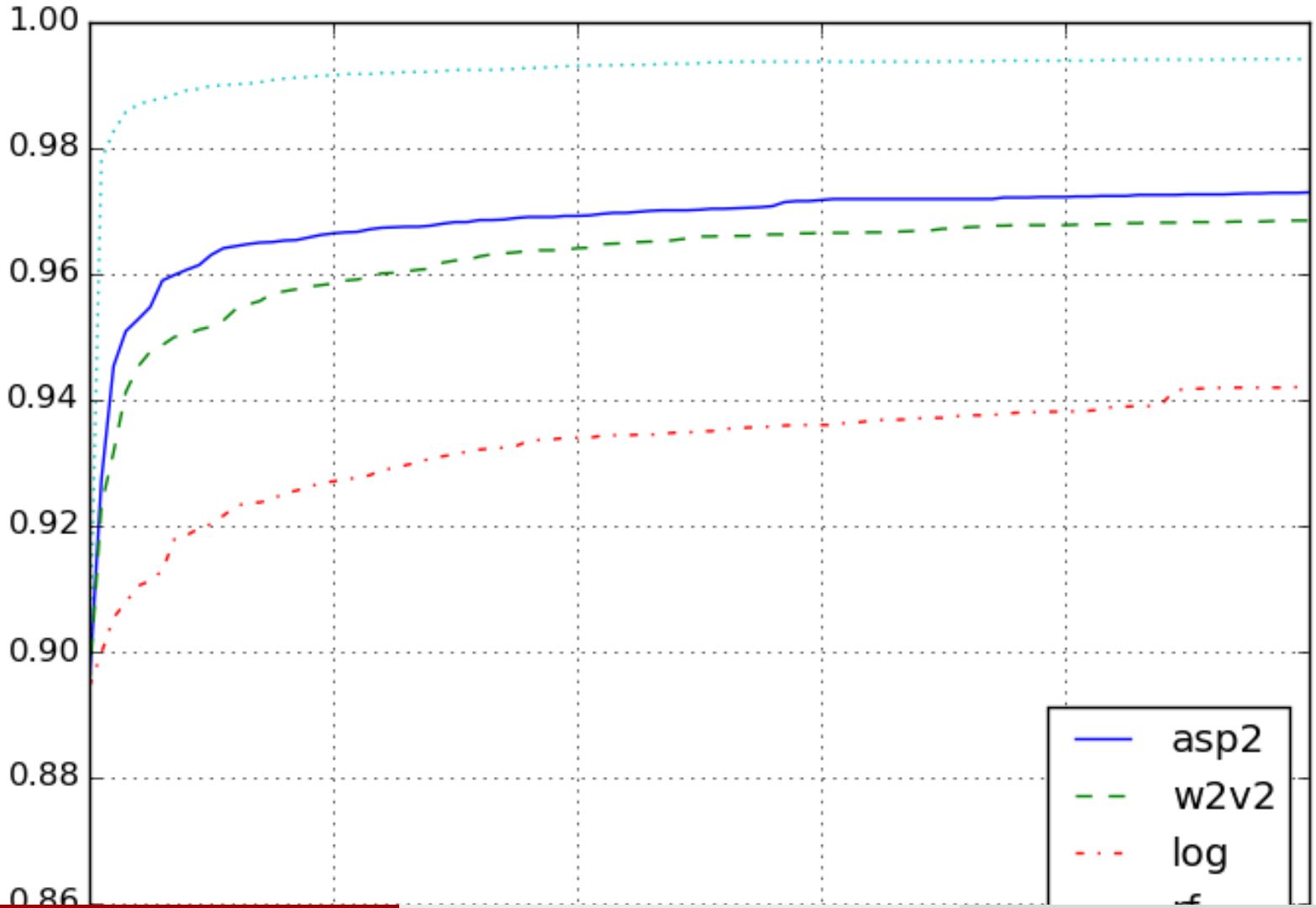


Taken from:

`databricks.com/blog/2015/01/21/random-forests-and-boosting-in-mllib.html`

Normalizatie: Ranking

Tweede poging: Random Forest Regressor

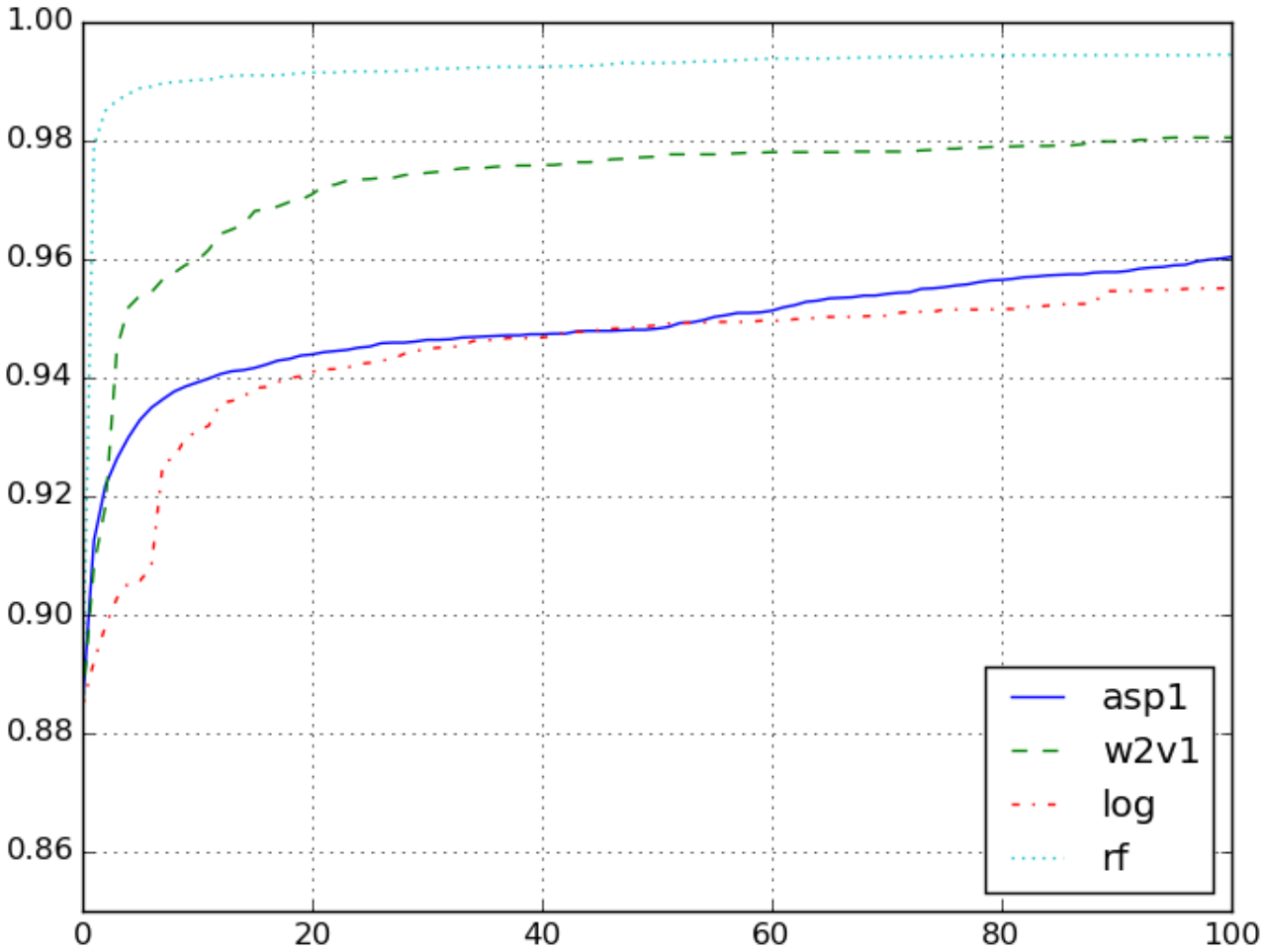


Normalizatie: Ranking

Tweede poging: Random Forest Regressor

Normalizatie: Ranking

Tweede poging: Random Forest Regressor



Normalizatie: Ranking

Tweede poging: Random Forest Regressor

Table of Contents

Parsen van Tweets: Earley Parser op basis van mogelijke inputs

Input = tabel

ur	da	boss
youre	ea	bos
your	ad	os
user		

Parsen van Tweets: Earley Parser op basis van mogelijke inputs

pos step

ur	0.1		
	NNP	0.8	0.08
	VB	0.2	0.02
youre	0.5		
	VB	1.0	0.5
your	0.2		
	PRP	0.8	0.16
	VB	0.2	0.04
user	0.2		
	NN	1.0	0.2
	1.0		1.0

$\mapsto$

NNP	0.08
VB	0.56
PRP	0.16
NN	0.20
	1.0

Parsen van Tweets: Earley Parser op basis van mogelijke inputs

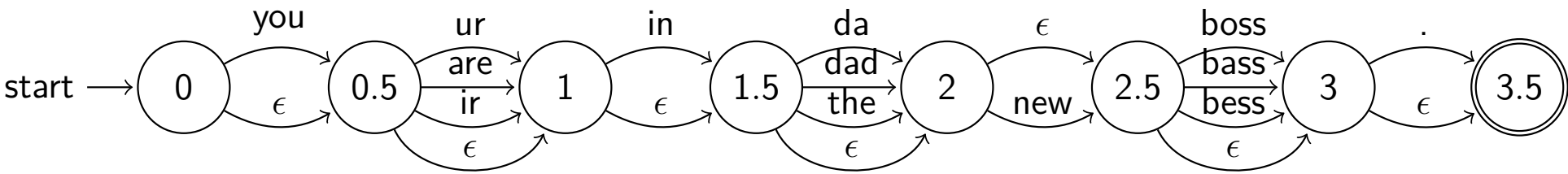
	Recall	Precision	F1	POS Accuracy
ark tagger	60.38	60.53	60.45	85.24
own tagger	61.37	61.35	61.36	85.04
stanford	69.43	68.37	68.90	83.33

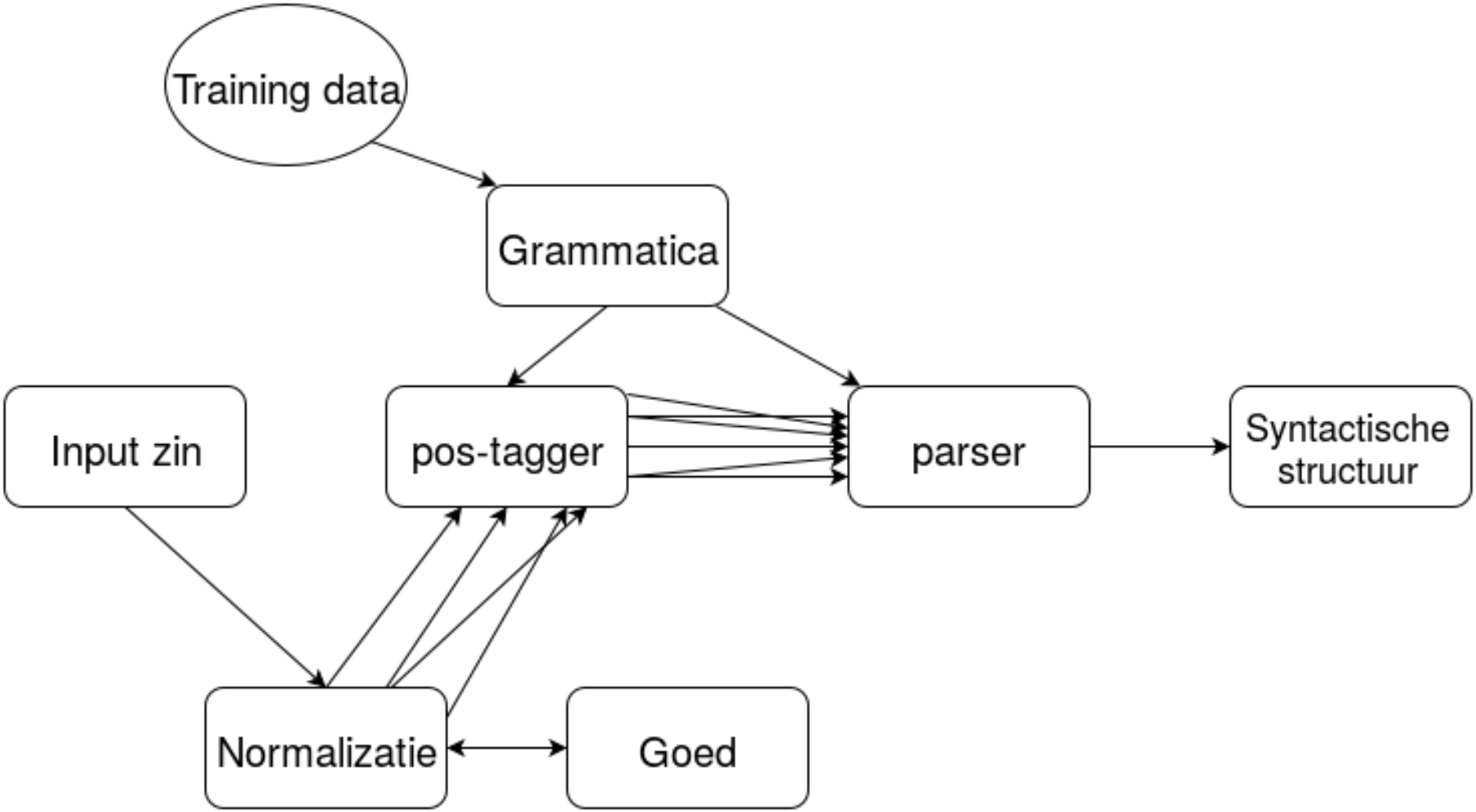
Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters Olutobi Owoputi, Brendan O Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith.


HOUSE OF THE BULL
WOK & ROLL
AUDIO SIGNAL PROCESSING
PHONO GRAFY
RECORD

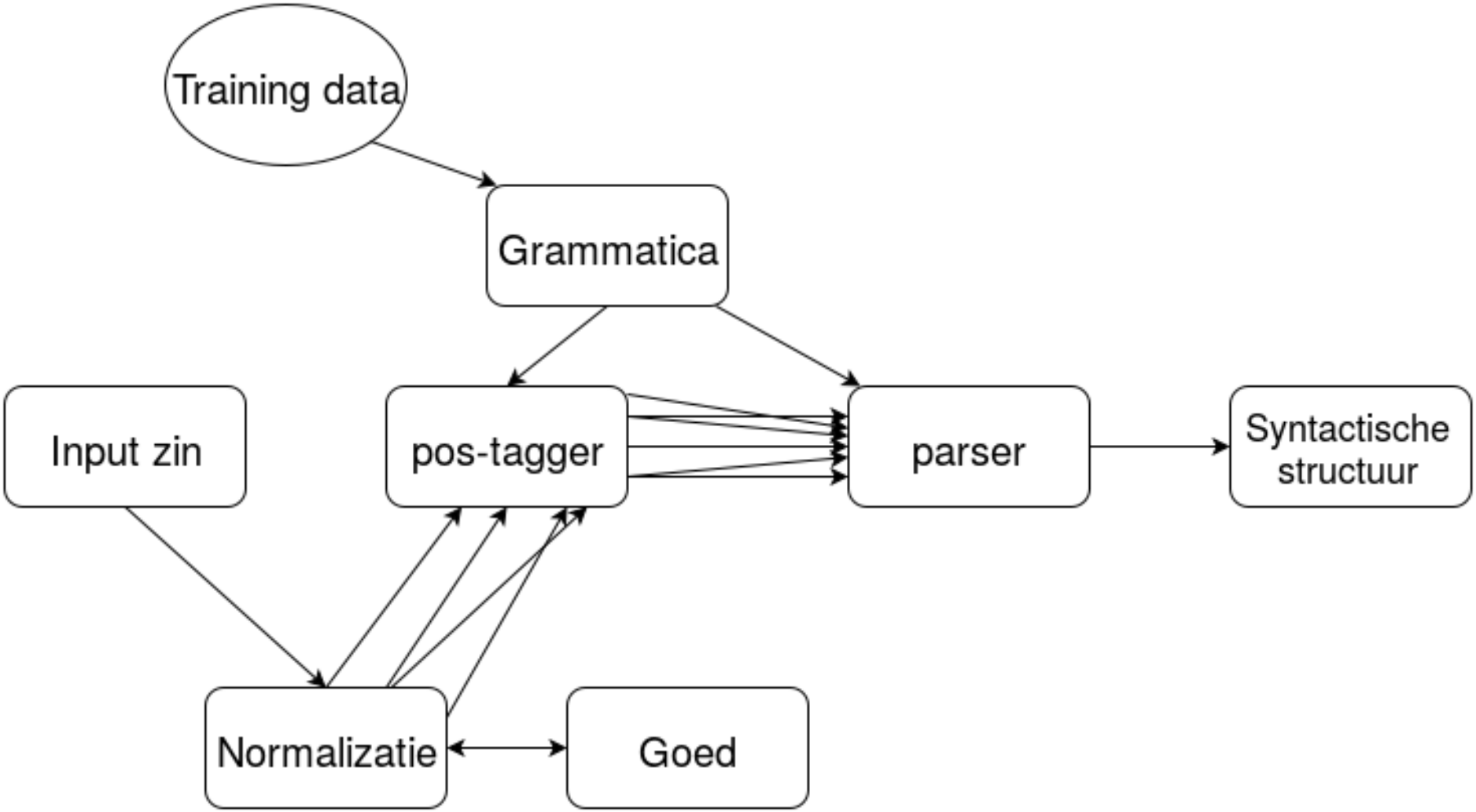
Parsen van Tweets: Parsing as Intersection

Input = fsa





- Van uitstel komt



- Van uitstel komt
- meer informatie

References

Roger Levy (2008): A noisy-channel model of rational human sentence comprehension under uncertain input.

Tyler Baldwin & Yunyao Li (2015): An In-depth Analysis of the Effect of Text Normalization in Social Media.

Chen Li & Yang Liu (2015): Joint POS Tagging and Text Normalization for Informal Text

Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn & Joseph Le Roux (2015): Foreebank: Syntactic Analysis of Customer Support Forums

Joachim Daiber & Rob van der Goot (to be published): The Denoised Web Treebank: Evaluating Dependency Parsing under Noisy Input Conditions

and: `http://siegfried.webhosting.rug.nl/~rob/`

Vragen?