Starten met NLP
in het Nederlands

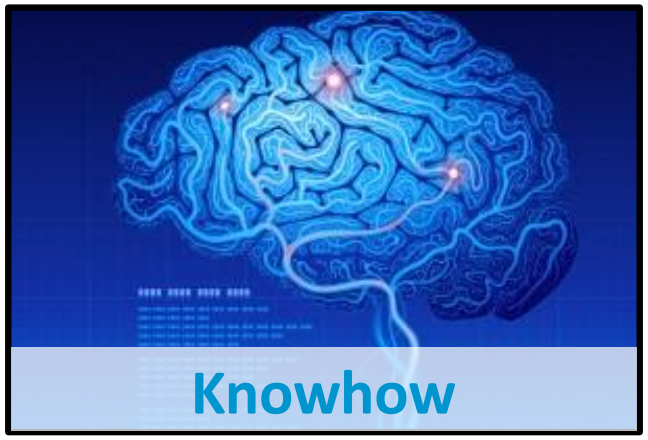
Joachim Ganseman
Smals Research

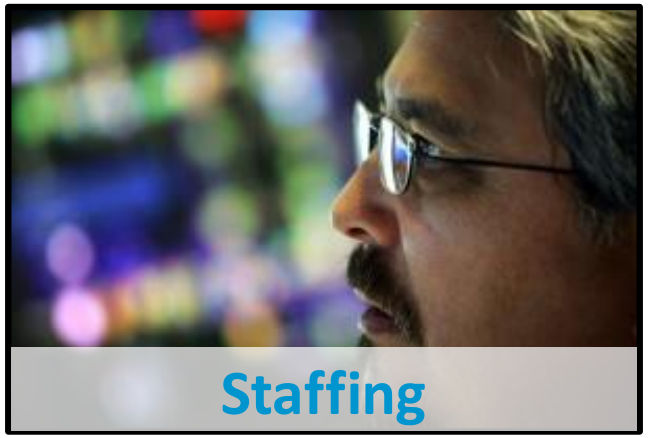
30/03/2021

Smals
ICT for society

Smals
ICT for society

SUPPORT FOR E-GOVERNMENT





g-cloud

top EMPLOYER BELGIË BELGIQUE
CERTIFIED EXCELLENCE IN EMPLOYEE CONDITIONS

WWW.SMALS.BE

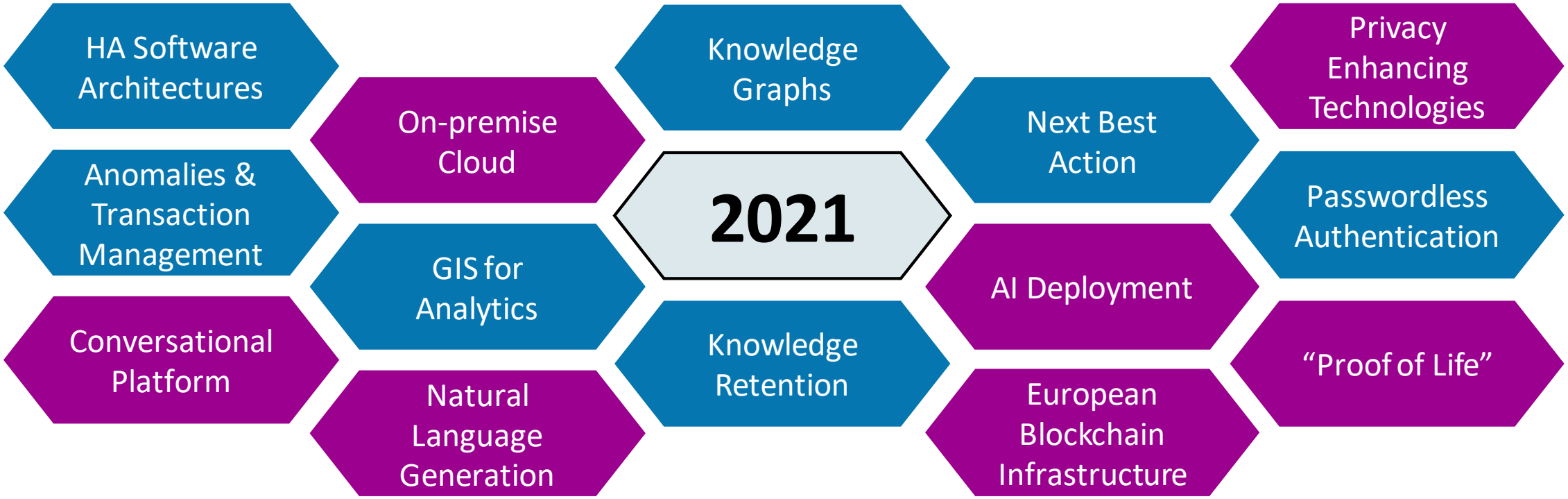
Smals Research 2021

Innovation with new technologies

Consultancy & expertise

Internal & external knowledge transfer

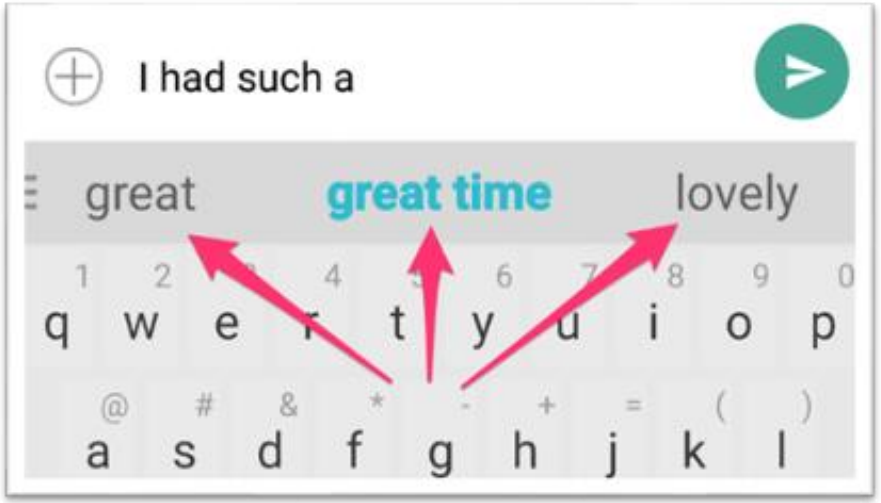
Support for going live



Smals ICT for society

www.smalsresearch.be

NLP: de computationele technieken waarmee we geschreven / gesproken tekst kunnen ontleden, analyseren en interpreteren.



Statistisch lettercombinaties associëren ≠ taal begrijpen!

NLP: subtaken

- Automatic speech recognition
- CCG
- Common sense
- Constituency parsing
- Coreference resolution
- Dependency parsing
- Dialogue
- Domain adaptation
- Entity linking
- Grammatical error correction
- Information extraction
- Language modeling
- Lexical normalization
- Machine translation
- Missing elements
- Multi-task learning
- Multi-modal
- Named entity recognition
- Natural language inference
- Part-of-speech tagging
- Question answering
- Relation prediction
- Relationship extraction
- Semantic textual similarity
- Semantic parsing
- Semantic role labeling
- Sentiment analysis
- Shallow syntax
- Simplification
- Intent Detection and Slot Filling
- Stance detection
- Summarization
- Taxonomy learning
- Temporal processing
- Text classification
- Word sense disambiguation

Current state-of-the-art: https://nlpprogress.com/

Smals
ICT for society

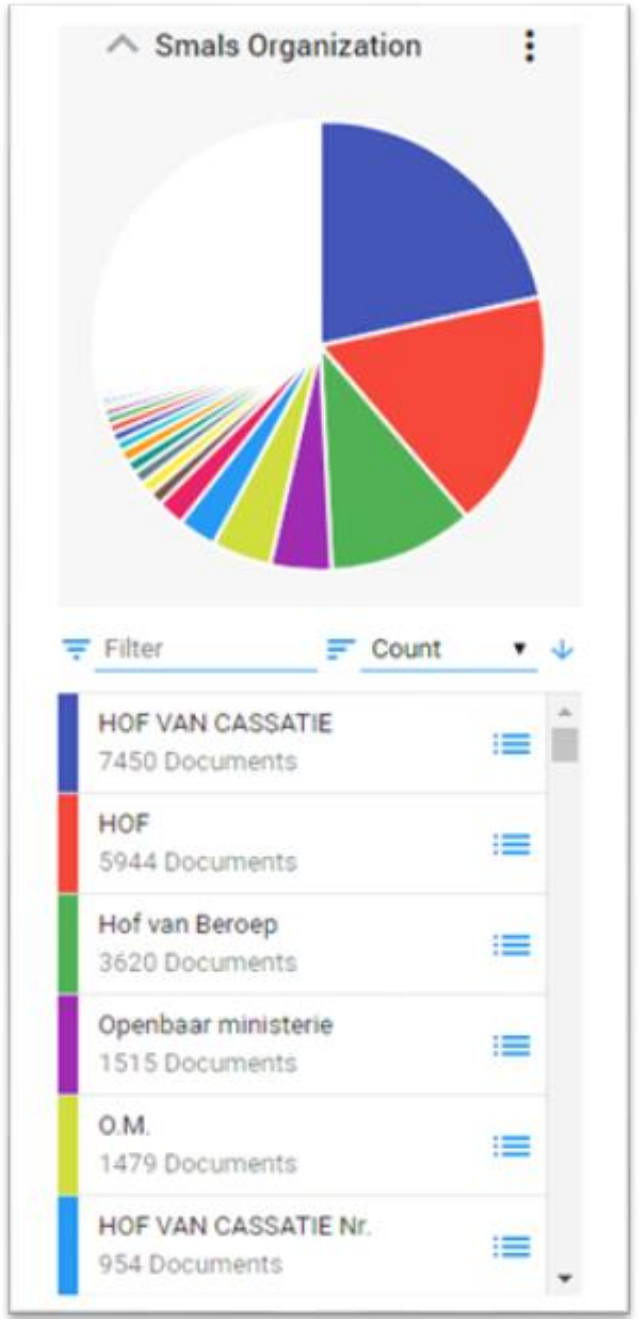
Named Entity Recognition

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

(voorbeeld © Europeana)



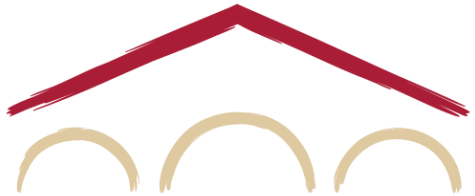
INTERVIEW

‘Alle Sky ECC-berichten lezen, zou ons 685 jaar kosten’

Het gerecht zou met de huidige middelen 685 jaar nodig hebben om alle berichten te lezen die onderschept zijn bij de kraak van de misdaadtelefoons van Sky ECC. Federaal procureur Frédéric Van Leeuw

(De Tijd, 12/02/2021)

Enkele NLP libraries



CoreNLP

AllenNLP

Stanza

flair

NLTK

Natural Language Toolkit

Reference Guide

spaCy

NLP ARCHITECT

GENSIM

topic modelling for humans

Smals

ICT for society

Een eerste blik

- ```
  import spacy
  from spacy import displacy
  nlp = spacy.load("en_core_web_sm")
  doc = nlp("Apple buys a French company for $1 billion.")
  displacy.render(doc, style="ent")
  ```

Apple ORG buys a French NORP company for $1 billion MONEY .

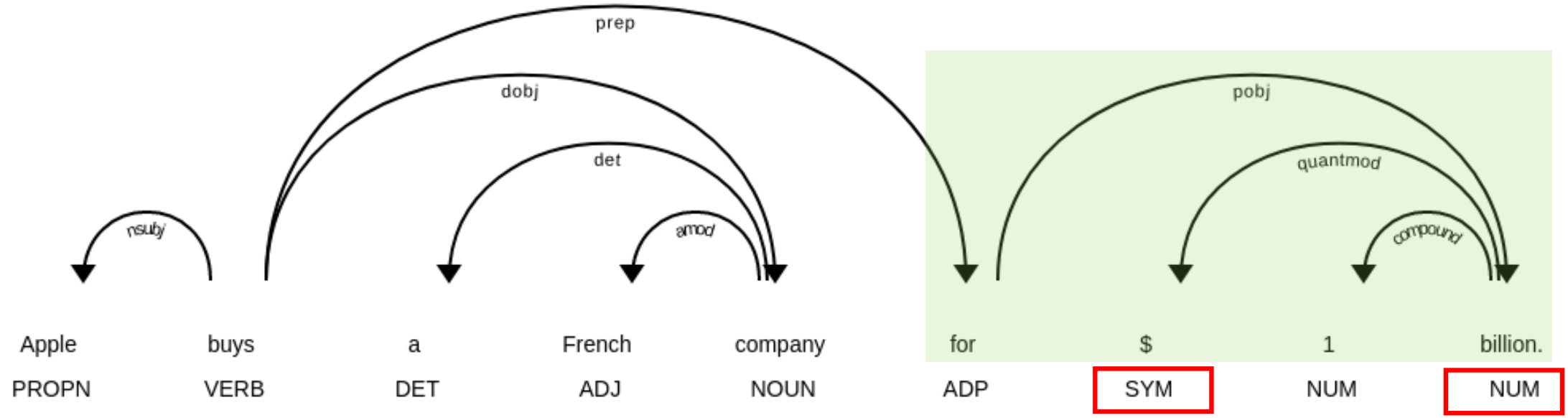
- ```
  nlp = spacy.load("nl_core_news_sm")
  doc = nlp("Apple koopt een Frans bedrijf voor $1 miljard.")
  displacy.render(doc, style="ent")
  ```

Apple PERSON koopt een Frans NORP bedrijf voor $ 1 CARDINAL miljard.

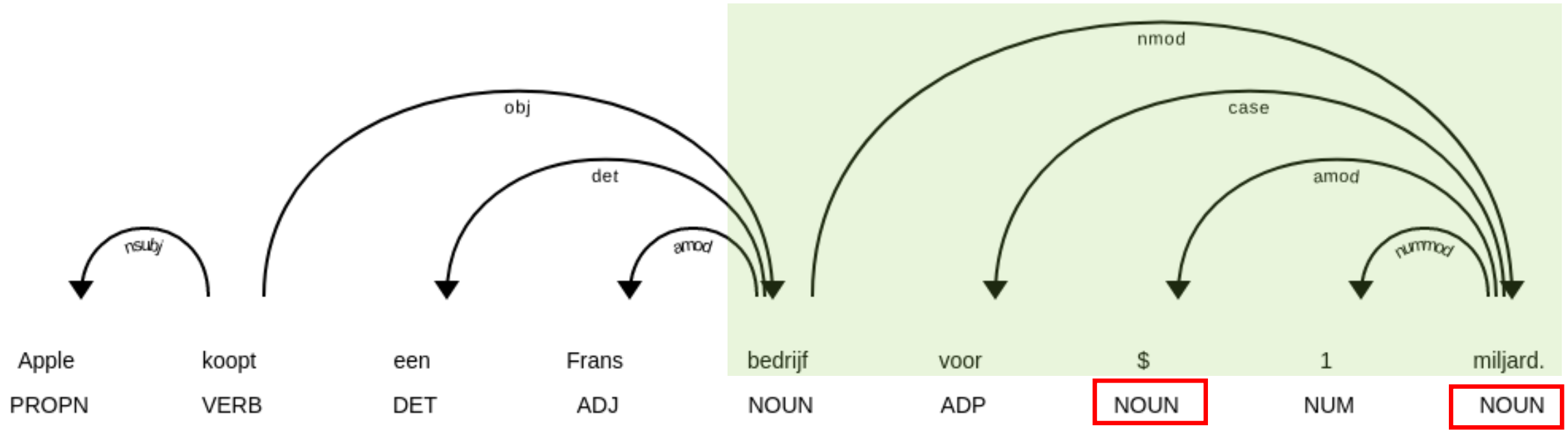
Smals
ICT for society

In de achtergrond

- EN:



- NL:



Verschillende talen, verschillende analyses

English

TAGGER

$, '' , , , -LRB- , -RRB- , . , : , ADD , AFX , CC , CD , DT , EX , FW , HYPH , IN , JJ , JJR , JJS , LS , MD , NFP , NN , NNP , NNPS , NNS , PDT , POS , PRP , PRP$, RB , RBR , RBS , RP , SYM , TO , UH , VB , VBD , VBG , VBN , VBP , VBZ , WDT , WP , WP$, WRB , XX , ``

49 woordsoort-labels

"a.m.",
"Adm.",
"Bros.",
"co.",
"Co.",
"Corp.",
"D.C.",

Een 100-tal afkortingen
(“tokenizer exceptions”)

Nederlands

N|soort|ev|basis|zijd|stan__Gender=Com|Number=Sing ,
N|soort|ev|dim|onz|stan__Gender=Neut|Number=Sing ,
N|soort|mv|basis__Number=Plur , N|soort|mv|dim__Number=Plur ,
SPEC|afgebr , SPEC|afk__Abbr=Yes , SPEC|deeleigen , SPEC|enof ,
SPEC|meta , SPEC|symb , SPEC|vreemd__Foreign=Yes , TSW ,
TW|hoofd|nom|mv-n|basis , TW|hoofd|nom|mv-n|dim ,
TW|hoofd|nom|zonder-n|basis , TW|hoofd|nom|zonder-n|dim ,
TW|hoofd|prenom|stan , TW|hoofd|vrij , TW|rang|nom|mv-n ,

… → 233 woordsoort-labels

"b.v.",
"b.ver.coll.gem.gem.comm.",
"b.verg.r.b.",
"b.versl.",
"b.wl.ev."

… → 1500+ afkortingen

Smals
ICT for society

Verschillen in trainingsdata

English

OntoNotes 5.0

- 300000 zinnen
- 2.9mln woorden
- Nieuws, wiki, fora, conversaties, bijbel, …

Nederlands

UD LassySmall + Alpino (*)

- 20966 zinnen
- 306764 woorden
- wiki + nieuws uit begin jaren ’00

(ref: Bijbelvertaling = ± 900000 woorden)

(* uitgez. word vectors: CommonCrawl+Wiki)

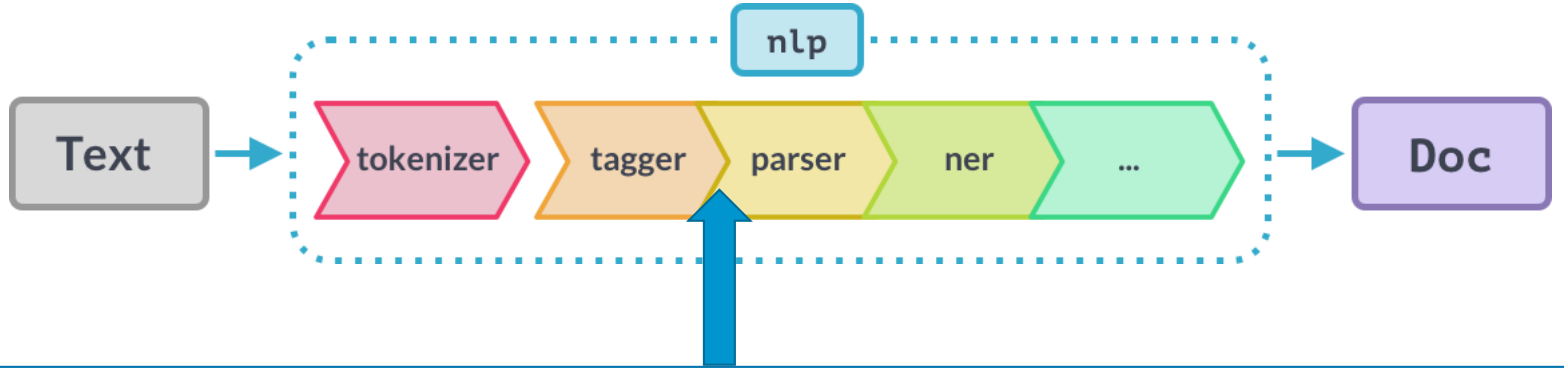
Model	Labeled dependencies	Part-Of-Speech (detailed)	NER F-score
Nl_core_news_sm (CNN)	81%	94%	72%
Nl_core_news_lg (CNN)	84%	95%	77%
En_core_web_sm (CNN)	90%	97%	84%
En_core_web_lg (CNN)	90%	97%	86%
En_core_web_trf (roBERTa)	94%	98%	90%

Situatie 25/03/2021

Smals ICT for society

De analyse verbeteren (1)

- Zelf gedefinieerde componenten invoegen in de analysepijplijn




```
ruler = nlp.add_pipe("attribute_ruler", name="fix_num", after="morphologizer")
detect = [[{"POS": "NOUN", "LIKE_NUM": True}]]
assign = {"POS": "NUM"}
ruler.add(patterns=detect, attrs=assign)
```

Apple	koopt	een	Frans	bedrijf	voor	$	1	miljard.
PROPN	VERB	DET	ADJ	NOUN	ADP	NOUN	NUM	NUM

- Goed voor specifieke verbeteringen, in functie van eindtoepassing

Smals
ICT for society

De analyse verbeteren (2): een component bijtrainen

- Kloon een van de voorbeeldprojecten (https://spacy.io/usage/projects)
  - `python -m spacy project clone pipelines/ner_demo_update`
- Pas de “makefile” *project.yml* aan
- Voeg eigen trainingsdata en preprocessing scripts toe:

```
["OnePlus 9 Pro met nieuwe Sony-sensor verschijnt eind maart voor 899 euro.",
        {"entities":[[0,7,"ORG"],[25,29,"ORG"],[64,72,"MONEY"]]}],
["Gerucht: Discord voert gesprekken met Microsoft over mogelijke overname.",
        {"entities":[[9,16,"ORG"],[38,47,"ORG"]]}],
...
```

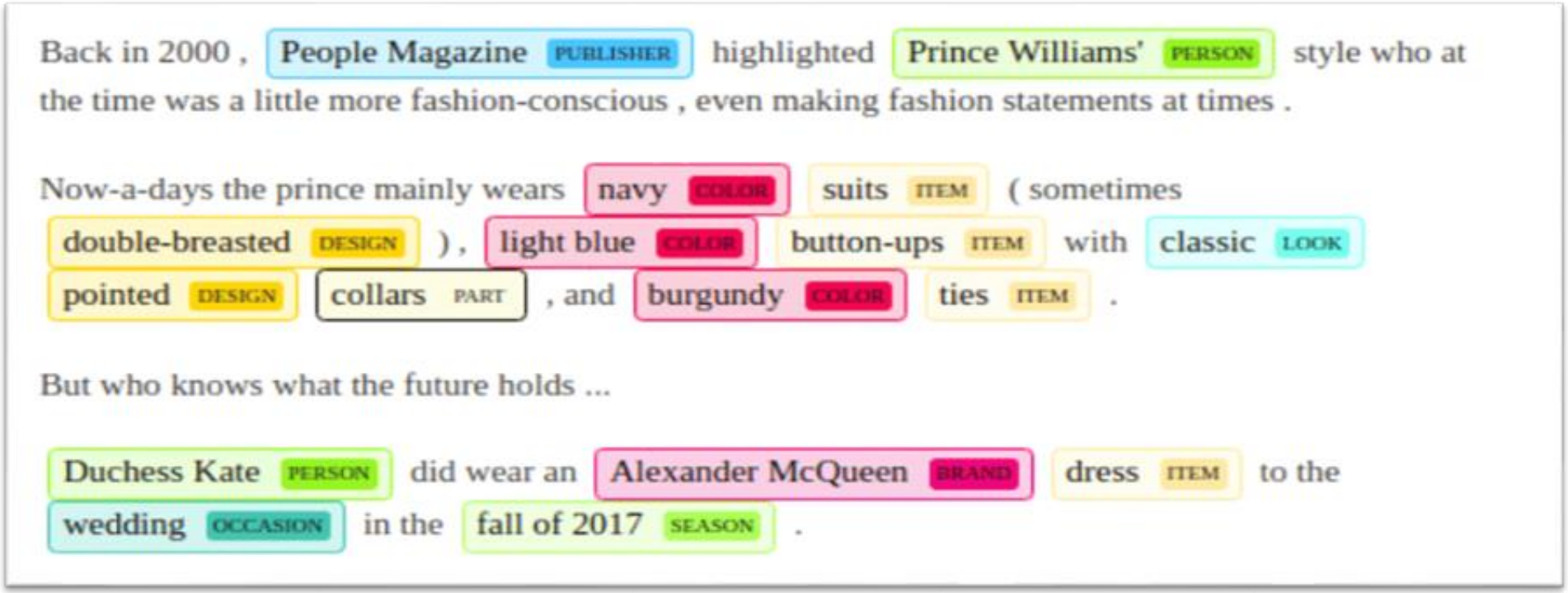
- `python -m spacy project run all`

- Risico: catastrophic forgetting → trainingsdata finetunen

Smals ICT for society

De analyse verbeteren (3) : een eigen NER trainen

- Vertrek ook hier van een bestaand project
  - `python -m spacy project clone pipelines/ner_demo_update`
- Leg een eigen, volledige dataset aan
- Behoudt alleen labels die je nodig hebt

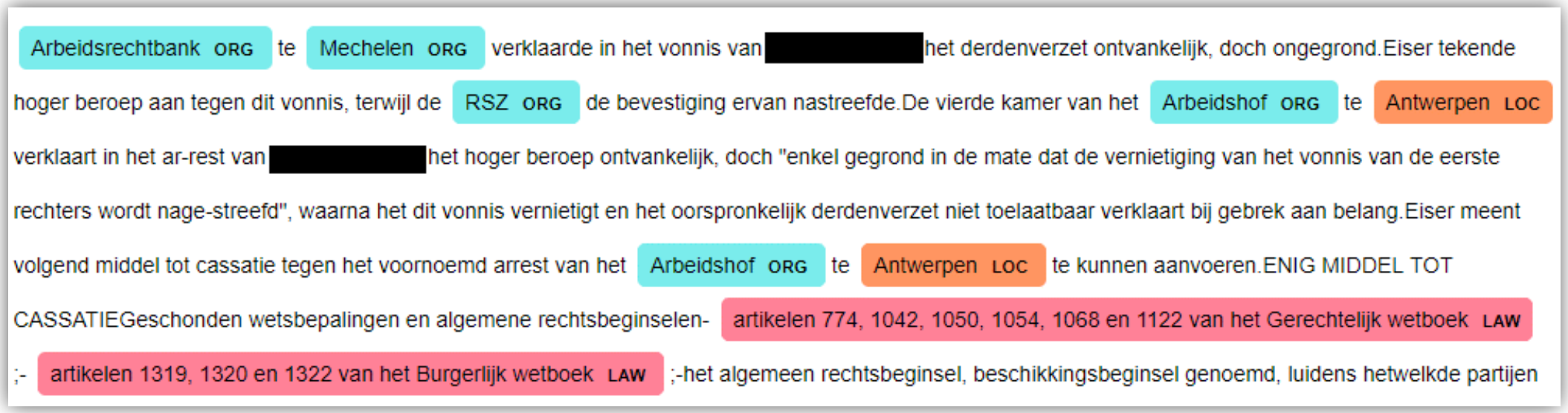


Smals
ICT for society

Toepassing: juridische tekst

- Dataset: gepubliceerde arresten Hof van Cassatie
- Train op entiteiten Locatie, Organisatie en Wetsartikelen
  - Maak dus een eigen geannoteerde trainingsdataset:

```
[ "toepassing van artikel 72 van de wet van 15 december 1980
betreffende de toegang tot het grondgebied",
      {'entities': [(15, 57, 'LAW')]}  ],
…
```



Smals
ICT for society

Een volgende stap: Entity Linking

- Bvb. herformatteer herkende entiteiten als URLs

  proof-of-concept uitgewerkt door TheMatchbox op NLP4Gov Hackathon, Informatie Vlaanderen, 2018:

- Met behulp van Knowledge Bases
  - Zie Sofie Van Landeghem, “Training a custom Entity Linking model with spaCy”, YouTube

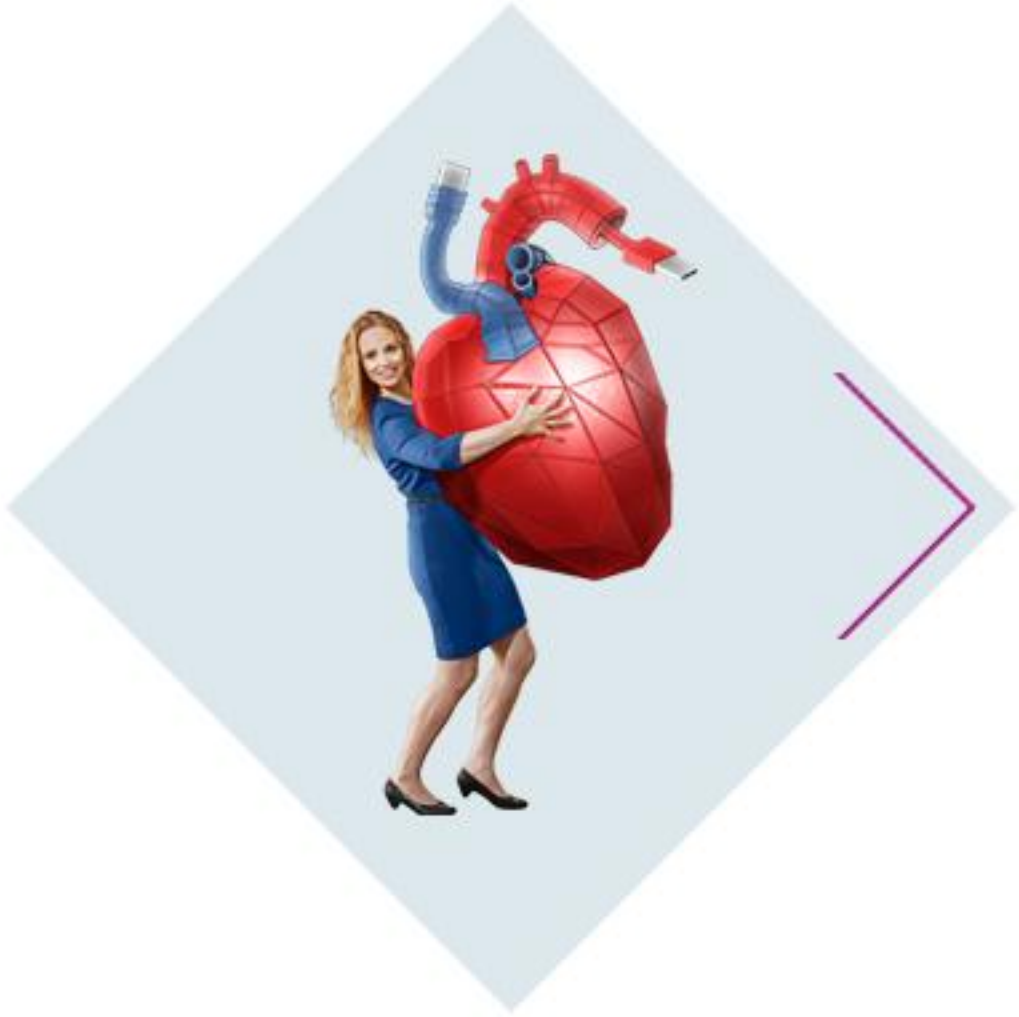
Smals
ICT for society

Conclusies

- Resterende problemen
  - Structureel gebrek aan grote Nederlandstalige geannoteerde datasets
  - Nederlandse taalmodellen lopen achter op Engelstalige

- Het goede nieuws
  - Zelf een taalmodel tweaken is gemakkelijk
    - ref. Wietse de Vries & Malvina Nissim, *"As good as new. How to successfully recycle English GPT-2 to make models for other languages"*, https://arxiv.org/pdf/2012.05628.pdf
  - Er komen in sneltempo Nederlandse taalmodellen bij
    - Zie o.a. https://huggingface.co/models : reeds 28 NL-talige transformer modellen
  - Grote interesse en funding vanuit EU
    - European Language Grid
    - CEF eTranslation
    - Streven naar "digital language equality": investeringen in *under-resourced languages*

Smals
ICT for society



Met dank aan

Katy Fokou
SpaCy devs & contributors
Sofie Van Landeghem
Yves Peirsman (NLP Town)
TheMatchbox
RU Groningen
…

www.smalsresearch.be

www.smals.be/jobs

Joachim Ganseman
joachim.ganseman@smals.be