

# MA-thesis – Text Simplification

Gertjan van Noord

November 22, 2019

# Text Simplification

- there are now various tools / websites / plug-ins which judge 'readability'
  - “Eenvoudig Nederlands” van Bureau Taal
  - “T-Scan” van Pander Maat ea
- next step: automatic text simplification?

# Predictors for Readability

Sentence length, word length, word frequency, embedded sentences, number of modifiers, use of personal pronouns, use of abstract nouns, use of passive, use of negation, dependency length . . . , . . .

# Text Simplification

Matthew Shardlow, A Survey of Automated Text Simplification. In: (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing 2014 <https://core.ac.uk/download/pdf/25778973.pdf>

“Simplicity is intuitively obvious, yet hard to define.”

“Text Simplification is a largely unsolved task.”

“Text Simplification is currently not a commercialised application.”

# Text Simplification: Approaches

- Lexical
- Syntactic
- Explanation Generation
- Statistical Machine Translation
  - “Simple Wikipedia” provides data for English
- Non-English Approaches

# Text Simplification: Challenges

- Resources
- Systems
  - for reader or for author?
- Techniques
  - need for word-sense disambiguation for lexical simplification
  - improve techniques for identifying candidates for simplification

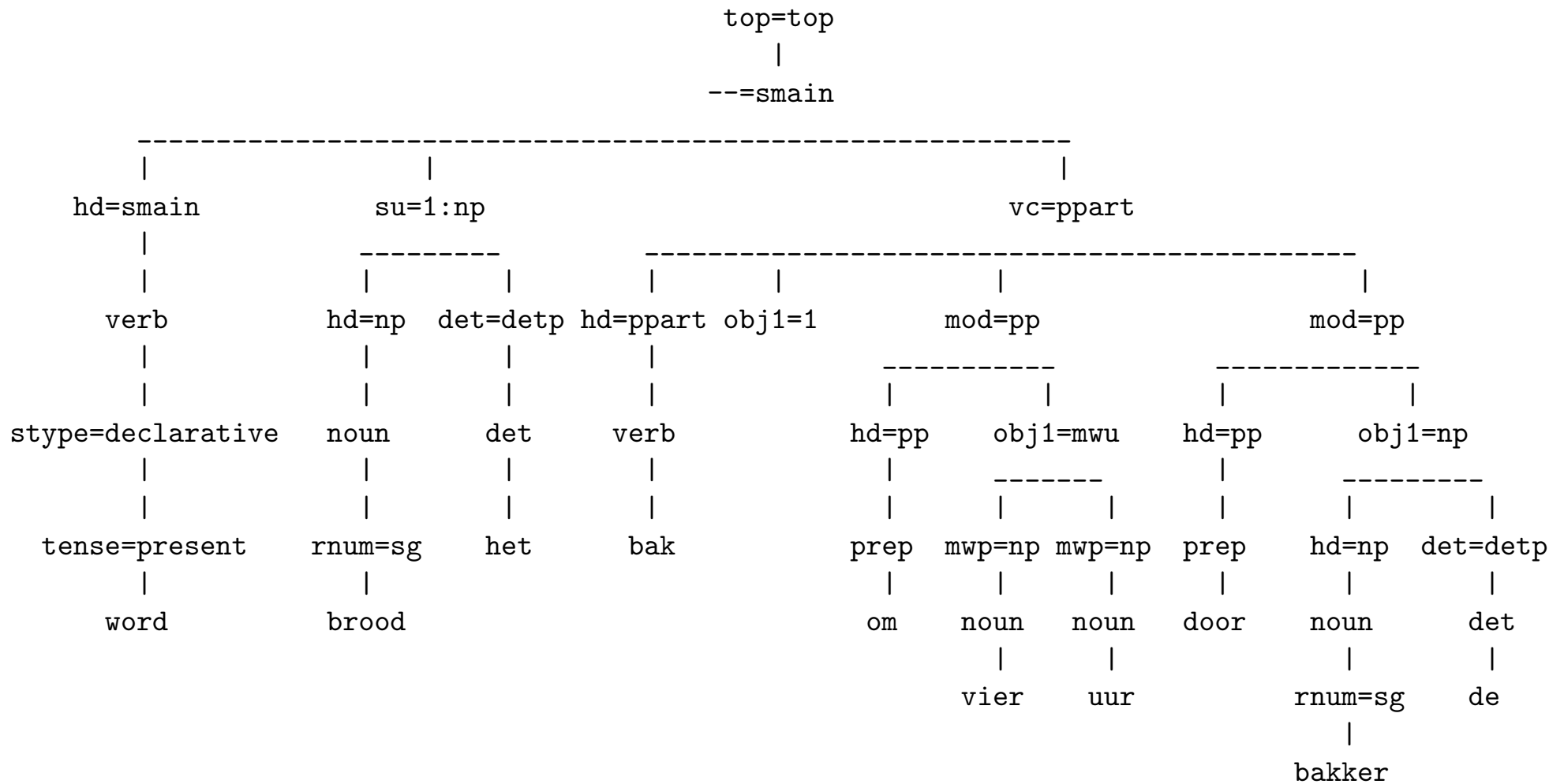
# Text Simplification for Dutch?

- rule based Text Simplification:
  1. Analyse the input into an abstract dependency representation
  2. Map the representation to a simpler representation
  3. Generate the output on the basis of this simpler abstract dependency representation
- Steps 1 and 3 already exist: Alpino

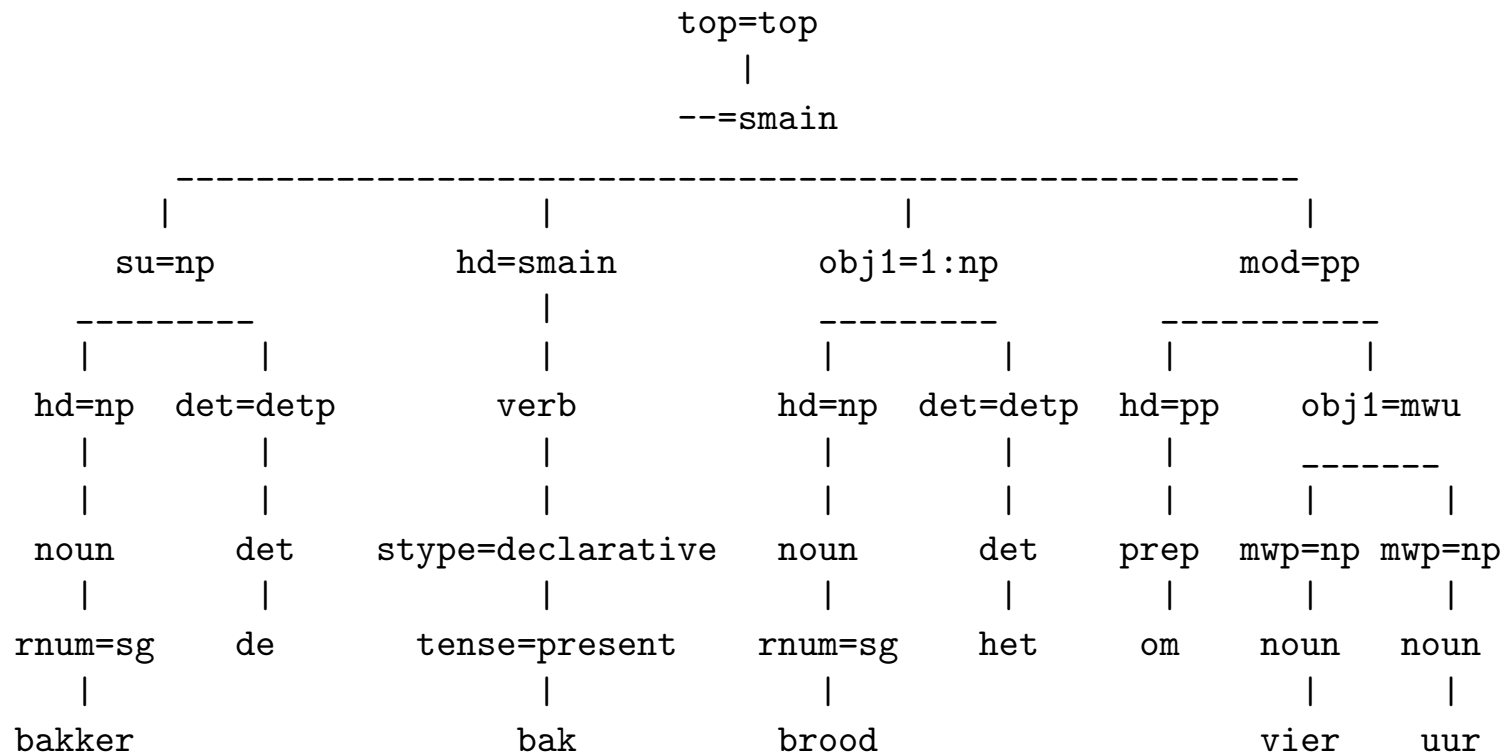
# Examples

- Passive to active transformation
- Remove modifiers
- Split conjunction of main clauses
- Split relative clause of subject of main clause
- Lexical simplification

# Het brood wordt om vier uur door de bakker gebakken



# Het brood wordt om vier uur door de bakker gebakken



# Open problems

This works very well. But . . .

## Open problems

This works very well. But: “wat je ook doet, de semantiek gooit roet” . . .

Spinnen worden voornamelijk gegeten door vogels

Vogels eten voornamelijk spinnen

De melkboer wordt graag door de buurvrouw verwend

De buurvrouw verwent graag de melkboer

de kinderen worden allemaal door een ouder opgehaald

Een ouder haalt de kinderen allemaal op .

De afgelopen tien jaar werden drie vrouwen uit het Haarlemse Blijfhuis door hun (ex-)man omgebracht

Hun (ex-)man bracht de afgelopen tien jaar drie vrouwen om uit het Haarlemse Blijfhuis

.

## Open problems

Not every door-PP is the agent . . .

Door haar publicatie in het Staatsblad is ze immers van kracht verklaard

Haar publicatie in het Staatsblad heeft ze immers van kracht verklaard

Die spoorbanen worden door druk bewoond gebied aangelegd

Druk bewoond gebied legt die spoorbanen aan

Op de US Open is gisteravond door de regen niet meer getennist

Op de US Open heeft de regen gisteravond niet meer getennist

Door de slechte zomer is er nauwelijks gebarbecued

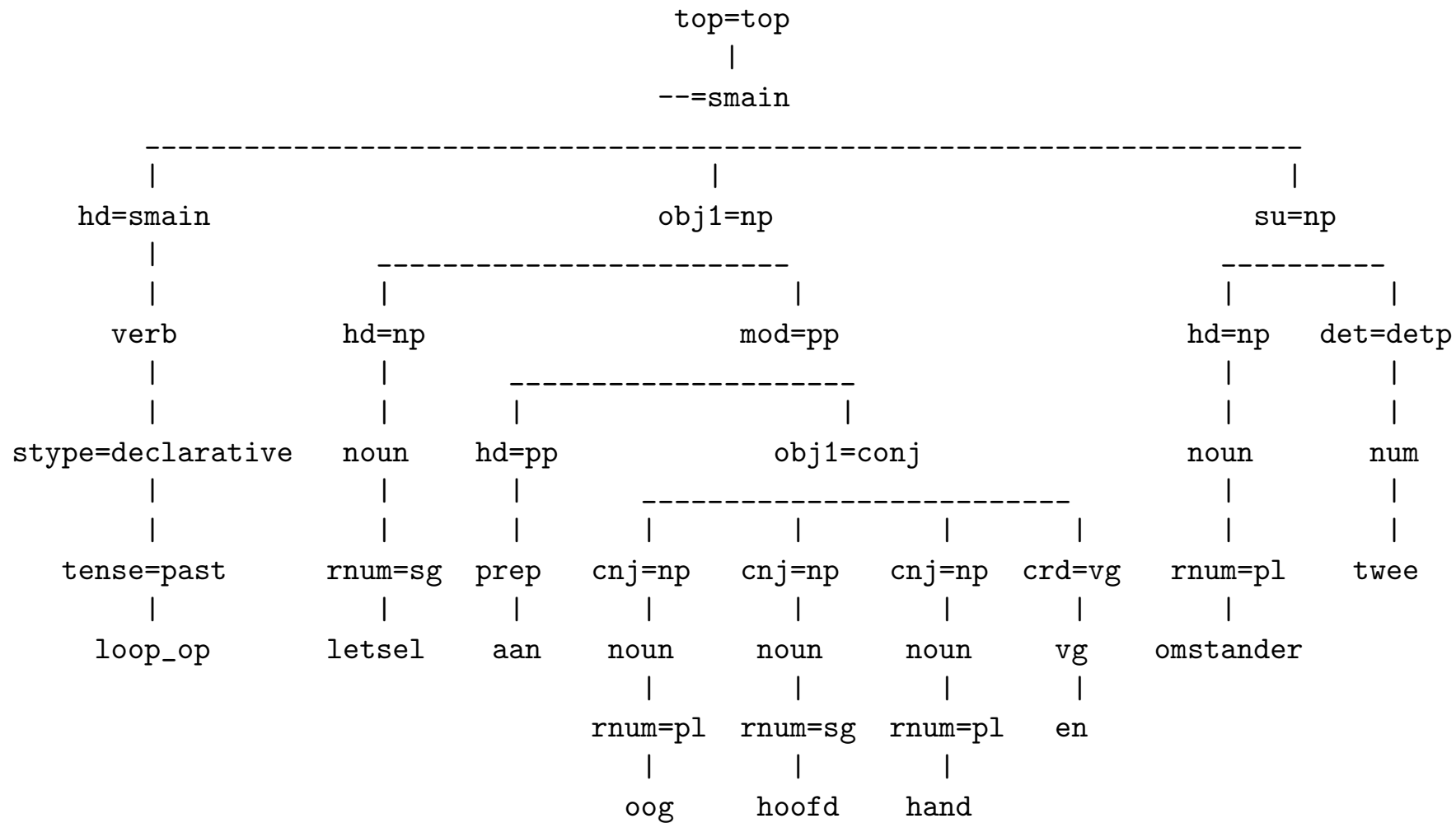
De slechte zomer heeft nauwelijks gebarbecued

## Remove modifiers

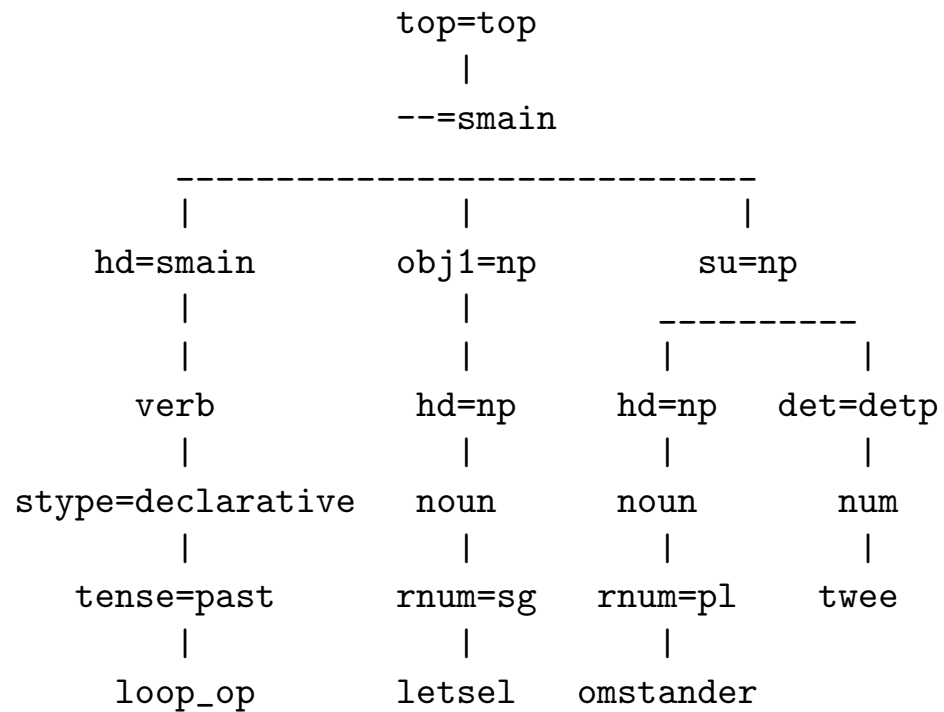
Obviously, this simplifies the text, but just as obviously, you cannot remove all modifiers in all situations . . .

“niet”, “nooit” are ruled out by means of a simple hack

# Twee omstanders liepen letsel op aan ogen , hoofd en handen



# Twee omstanders liepen letsel op



# Examples

Milieuvervuiling kost in Nederland net zoveel mensenlevens als ongezond leven of verkeersongevallen .  
Milieuvervuiling kost zoveel mensenlevens als leven of verkeersongevallen .

Volgens het Natuur en Milieuplanbureau gaan per jaar 18.000 mensen 10 jaar eerder dood door met name luchtvervuiling .  
18.000 mensen gaan dood .

Fijnstof is de grote boosdoener , zo blijkt uit de jaarlijkse Milieubalans , die vandaag werd gepresenteerd .  
Fijnstof is de boosdoener , zo blijkt uit de Milieubalans .

De huisartsen gaan opnieuw actie voeren tegen de nieuwe tarieven .  
De huisartsen gaan actievoeren .

De huisartsen voelen zich door minister Hoogervorst bedrogen .  
De huisartsen voelen zich bedrogen .

Zorgverzekeraars krijgen namelijk zeggenschap over een deel van het geld dat nu nog direct naar de huisartsen gaat .  
Zorgverzekeraars krijgen zeggenschap .

Ze zijn bang dat ze d'r heel erg op achteruit gaan .  
Ze zijn bang dat ze d'r op achteruitgaan .

## More examples

In Iran heeft de invloedrijke geestelijke H. Rafsanjani bekendgemaakt dat ie zich kandidaat stelt voor de presidentsverkiezingen in juni

H. Rafsanjani heeft bekendgemaakt dat hij zich kandidaat stelt voor de presidentsverkiezingen

Tussen 1989 en 1997 was ie ook al president van Iran .

Hij was president .

De gematigd conservatieve Rafsanjani wil betere betrekkingen met het westen .

De Rafsanjani wil betrekkingen .

Volgens de peilingen maakt ie een goede kans opnieuw president te worden .

Hij maakt een kans president te worden .

Unicef zegt dat 300.000 kinderen in Ethiopie nog dit jaar zullen sterven aan ondervoeding , als er geen hulp komt .

Unicef zegt dat 300.000 kinderen zullen sterven aan ondervoeding .

Het eind van een tijdperk .

Het eind .

Er zijn problemen.

Problemen zijn.

Daar werd toen volop gevoetbald.

*no result, because “gevoetbald werd” is not a sentence*

## Split conjunction of two main clauses in separate sentences

Redbald vergat de belofte van de doop en na enkele dagen hoorden de Friezen dat hij gestorven was .  
Redbald vergat de belofte van de doop . Na enkele dagen hoorden de Friezen dat hij gestorven was .

De bouwstijl is een traditionele Almohadenstijl en de toren is voorzien van vier koperen globes .  
De bouwstijl is een traditionele Almohadenstijl . De toren is voorzien van vier koperen globes .

Qua grootte was de piramide ongeveer gelijk aan die van Cestius en hij was eveneens met wit marmer bekleed .  
Qua grootte was de piramide ongeveer gelijk aan die van Cestius . Hij was eveneens met wit marmer bekleed .

## Split conjunction of two main clauses in separate sentences

Deze laatste kolonies groeien van jaar tot jaar en het lijkt erop dat de dieren er zich permanent gevestigd hebben .

Van jaar tot jaar groeien deze laatste kolonies . Het lijkt erop dat de dieren zich er permanent hebben gevestigd .

De Amerikanen werden hierdoor verrast , en diplomatiek en militair ingrijpen bleef niet uit .

Hierdoor werden de Amerikanen verrast en diplomatiek . Militair ingrijpen bleef niet uit .

Men schrok erg en plannen werden gemaakt om Nederland weer leefbaar te maken voor de otter .

Men schrok erg . Plannen werden gemaakt om Nederland weer leefbaar te maken voor de otter .

## Generate pronoun

Deze voertuigen kunnen maximaal 2 passagiers vervoeren en zijn erg snel en wendbaar .

Deze voertuigen kunnen maximaal 2 passagiers vervoeren . *Deze voertuigen* zijn erg snel en wendbaar .

Deze voertuigen kunnen maximaal 2 passagiers vervoeren . *Ze* zijn erg snel en wendbaar .

Deze voertuigen die zo'n 12000 euro moeten kosten kunnen maximaal 2 passagiers vervoeren en zijn erg snel en wendbaar .

Deze voertuigen die zo'n 12000 euro moeten kosten kunnen maximaal 2 passagiers vervoeren . *Deze voertuigen die zo'n 12000 euro moeten kosten* zijn erg snel en wendbaar .

Deze voertuigen die zo'n 12000 euro moeten kosten kunnen maximaal 2 passagiers vervoeren . *Ze* zijn erg snel en wendbaar .

## Split conjunction of two main clauses in separate sentences, with generated pronoun

De toren is voorzien van op elkaar gestelde steunberen en heeft een polygonale traptoren aan de zuidzijde .

De toren is voorzien van steunberen gesteld op elkaar . Ze heeft een polygonale traptoren aan de zuidzijde .

Ajax kreeg twee rode kaarten en verspeelde de voorsprong

Ajax kreeg twee rode kaarten . Het verspeelde de voorsprong .

De meest in het oog springende daarvan staan links op zijn buik en zijn portretten van Willy Alberti en Johnny Jordaan .

De meeste in het oog springende daarvan staan links op zijn buik . Ze zijn portretten van Willy Alberti en Johnny Jordaan .

## Generate pronoun – also useful for other transformations?

Deze voertuigen die zo'n 12000 euro moeten kosten kunnen maximaal 2 passagiers vervoeren  
*Deze voertuigen* kunnen maximaal 2 passagiers vervoeren . *Ze* moeten zo'n 12000 euro kosten .

In 1942 werd een Brusselse postzegelhandelaar , die in het bezit was van twee exemplaren , vermoord .  
In 1942 werd *een Brusselse postzegelhandelaar* vermoord . *Hij* was in het bezit van twee exemplaren .

Over zijn rug liepen benige knobbels , die als bescherming werkten .  
*Benige knobbels* liepen over zijn rug . *Ze* werkten als bescherming .

Daartegenover zit het kamerorkest , dat een constante achtergrond ( en echo ) verzorgt .  
Daartegenover zit *het kamerorkest* . *Het* verzorgt een constante achtergrond en echo .

Door het zuurstoftekort dat zo ontstaat , komen er stofjes vrij die een ontsteking in gang zetten  
Door het zuurstoftekort dat zo ontstaat komen er *stofjes* vrij. *Ze* zetten een ontsteking in gang .

# Split relatives

You have to recognize the type of relative clauses which allows this (restrictive versus non-restrictive relative clause).

Het aantal kastelen dat gebouwd werd in de middeleeuwen is groot .

Het aantal kastelen is groot. Ze werden gebouwd in de middeleeuwen .

Ook de portretten die Avedon maakte zijn indrukwekkend .

Ook de portretten zijn indrukwekkend . Avedon maakte ze .

De liedjes die werden gezongen werden beter ontvangen .

De liedjes werden beter ontvangen . Ze werden gezongen .

De functionele groepen die hierbij horen zijn benzylbromide en een cyanideanion

De functionele groepen zijn benzylbromide en een cyanideanion . Ze horen hierbij .

## Split relative clauses and conjunctions

Muren , plafonds en de vloer zijn glanzend geslepen en er staan tekens die de kleermaker niet kent .  
Muren , plafonds en de vloer is glanzend geslepen . Er staan tekens . De kleermaker kent ze niet .

Er komen drie kleermakers bijeen en de derde is een spring-in-het-veld die nergens voor deugt .  
Drie kleermakers komen er bijeen . De derde is een spring-in-het-veld . Dat deugt nergens voor .

Ze gaan naar het veld en er komt een oude vrouw die vraagt wat er aan de hand is .  
Ze gaan naar het veld . Er komt een oude vrouw . Ze vraagt wat er aan de hand is .

De Franken willen de heidenen in het noorden bereiken en de koning , die leeft bij Thonar en Balder , wil ze breken .  
De Franken willen de heidenen in het noorden bereiken . De koning wil ze breken . Hij leeft bij Thonar en Balder .

De postzegel , die het stadhuis van Dendermonde afbeeldt , werd uitgegeven in 1920 en was verkrijgbaar vanaf 5 augustus .  
In 1920 werd de postzegel uitgegeven . Hij beeldt het stadhuis van Dendermonde af . Vanaf 5 augustus was de postzegel verkrijgbaar .

# Replace difficult word with simpler synonym

Relatively straightforward, but you need:

- collection of synonyms
- perhaps word sense disambiguation
- notion of simplicity

# Potential projects

Ideas:

- Passive: is door-PP the agent? Classification task?
- Modifiers: which modifiers can be deleted? Classification task?
- Split relatives: classify restrictive/non-restrictive?
- Simpler synonyms: generate synonyms? Determine simplicity?

Worries:

- Evaluation

## Try it out yourself

```
export ALPINO_HOME=/net/aps/haytabo/src/Alpino  
export PATH=$ALPINO_HOME/bin:$PATH
```

```
Alpino -notk -flag paraphrase_file src/passive.pl
```

```
Alpino -notk -flag paraphrase_file src/modifier.pl
```

```
Alpino -notk -flag paraphrase_file src/split.pl
```

```
Alpino -notk -flag paraphrase_file src/simple_words.pl
```

and then:

```
para Dit is een zin
```

## Example

```
$ Alpino -notk demo=off debug=0 batch_command=paraphrase
using Alpino from directory /net/aistaff/vannoord/z/Alpino ...
% restoring /dev/shm/Alpino.sav...
% /dev/shm/Alpino.sav restored in 270 msec 145931904 bytes
hdrug: process 18798 on host haytabo (datetime(2019,9,11,10,2,57))
0 para> Het brood wordt door de bakker gebakken
% compiling /net/aistaff/vannoord/z/Alpino/src/passive.pl...
% module alpino_user_transformation imported into user
% compiled /net/aistaff/vannoord/z/Alpino/src/passive.pl in module alpino_
[Het brood wordt door de bakker gebakken]
Q#1|Het brood wordt door de bakker gebakken|1|1|-10.336150265500002
De bakker bakt het brood .
1 para> Het brood wordt al om vier uur gebakken
[Het brood wordt al om vier uur gebakken]
Q#2|Het brood wordt al om vier uur gebakken|1|1|-10.190335962940003
Het brood bakt men al om vier uur .
```