TELECATS

Radboud University

Spreek2Schrijf

(Speak2Write)

An attempt to go from
spoken audio to the written Hansard

UNIVERSITY OF TWENTE.

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

TELECATS

Disclosure of spoken documents

Arjan van Hessen

TELECATS

Why here?

We do the automatic subtitling of the plenary session of the Tweede Kamer

We got a grant of the TK for the development of S2S

Speaker

Government

Presidium

Interruptions & questions

What can we do with HLT to support the Reporting Office in their main task: making the Dutch Hansard?

TELECATS

Why using HLT?

Members of the public and management say:

"why don't you use ASR for speech-to-text?"

Faster

A 00:05:49 fragment was processed in 2 minutes

Cheaper

If you replace the humans by computers yes, but….

Better

No, not yet. But from spoken to written…..

Extra possibilities

Research

Retrieve

Search

Speaking styles

Emotions

Semantics

Is it already used?

TELECATS

Already used for some years

Everything that is spoken is recorded with multiple cameras and microphones

The Hansard text

Human made transcriptions

Realign the correct text with the audio

Generate subtitles & Karaoke style text

Does it help the reporters?

No, it only adds new functionalities

Helps scholars with their research

It shows a possible future use of HLT

TELECATS

Spoken2Written

Develop an algorithm that “learns” to transcode ASR-output into text that is close to the written version of the spoken utterances

TELECATS

S2S – Main goal of the project

Develop a system that helps the “Reporting Office” to make the written Hansard.

Verbatim transcribe the spoken speech with ASR.

Clean the ASR-results (ehs, hmm), compound the decompound words.

Transcode these results into a written-alike version

Provide the recognised speech and an text editor to the transcriber for correction.

Realign the corrected text with the audio.

Generate subtitles & Karaoke style text.

Generate the Hansard text.

How will we do that?

TELECATS

S2S – Building

Improve the Automatic Speech Recognition results by using a better/more dedicated Parliamentary Language Model

A new LM has been build, based om 10 years of Hansard data

Make it possible to add “meeting specific” words **before** the recognition

More or less possible, but not user friendly

Change the output (a string of words) into phrases.

Statistical and pronunciation approach

Next month start with training **Moses**: the Open source version of Google Translate for this purpose

TELECATS

Building S2S

Build a Parliamentary Language Model (PML)

Collect 200M words from the last years

Recognise 100 hours of Plenary speech

Collect the Hansard text of these 100 hours

L1 (S)

Uses Moses to translate spoken into written

L2 (W)

Parallel Corpora

MOSES
statistical
machine translation
system

Moses

Overview
Manual
Online Demos
FAQ
Mailing Lists
Get Involved
Recent Changes

Getting Started

Source Installation
Baseline System
Packages
Releases
Sample Data

Welcome to Moses!

Moses is a **statistical machine translation system** that allows you to automatically train translation models for any language pair. All you need is a collection of translated texts (parallel corpus). Once you have a trained model, an efficient search algorithm quickly finds the highest probability translation among the exponential number of choices.

News

- **8 September 2016** Moses2, a fast drop-in replacement for the Moses decoder
- **12 December 2015** Add a new feature function to Moses
- **17 June 2015** Slate for Windows
- **15 June 2015** Moses, and more, on Amazon cloud Box
- **1 June 2015** Developing Moses with Eclipse **video**
- **4 February 2015** Moses v 3.0 has been **released!**
- **21 July 2014** Moses now has nightly **speed tests**
- **14 July 2014** **How to compile Moses with Eclipse**

Moses on Twitter

Tweets by @MosesSMT

Moses SMT
@MosesSMT

Sign up for the MT Marathon in Lisbon. Special focus on NMT. Great speakers and great food. mtm2017.unbabel.com

Jul 14, 2017

Moses SMT Retweeted

Barry Haddow
@bazril

Results rolling in for #wmt2017 news

Problems

Dutch is (like German) a compound language

Difference between spoken speech and written text by humans

ASR output is a string of words and not a phrase

TELECATS

Building the Parliamentary Language Model

200M words used for building the PLM

750.000 unique words

A “normal” LM contains 250.000 unique words

TELECATS

S2S – Decompounding - Compounding

Decompound these words before building the Language Model

Compound these words after the recognition

Short-term negotiation goals → Kortetermijnonderhandelingsdoelstellingen

Environmental reporting report → Milieurapportagerapport

Gekwalificeerdemeerderheidsbesluitvorming

Internationalebedrijfsleveninstrumentariumin

Vorderingsvrijstellingsovergangsmaatregel

Langeretermijnhoudbaarheidsdoelstellingen

Ontwikkelingssamenwerkingsverantwoordelijkheid

S2S – (No-)Phrases

Will see if Moses can do this

Will see if pauses and some heuristics can (partly) deal with this problem

TELECATS

Differences between spoken and written text

Well, I want to stress that, eh that, well because …..

I want to stress that because …..

Human imperfection

It was warm, so I drunk a beer

I drunk a beer because it was warm

Human pursuit of perfection

However, this seems to change as Wouter Zwijnenburg promised last Sunday during his talk.

What do we hope to achieve?

Perfection?

No humans involved?

TELECATS

What do we hope to achieve?

A system that produces reasonable good results

That helps the Hansard writers to:

- Speed-up their work
- Concentrate on the less boring parts

Automatically generate subtitles based on ASR

Offer other possibilities for text presentation (*Karaoke in stead of subtitles)*

Increase the search performance (topic search)

Increase the online access (*fast and reliable*) to:

- Everything spoken in Parliament
- Offer access to the searched spoken fragments (*both AV and text*)
- Automatic generated summaries of each item
- Debate graphics / analysis

TELECATS

Ik ben het met die beesten gaan we echt niet in één jaar.

Die toch zit hier het problematiek hebben. Roemer begint met dat we het eens moeten zijn over dat het systeem niet werkt. Nou ik kan u zeggen. Niks is zo bureaucratisch als beginnen met het veranderen van systemen. Dus dat is gewoon niet verstandig als het werk. Nogmaals , gaan we naar kijken maar doet het niet eerder deze week zag een foto van u in hemdsmouwen over de zorg met een cirkelzaag. Dit is dit is die cirkelzaag. Maar het werkt niet. In Engeland de en eet je een niet te betalen systeem met oudere en slechtere ziekenhuis dan we hier hebben met hoge kosten en heel veel privéklinieken voor de rijken. Ik kan het niet mooier maken. Het is het is niet dat ik het wil. Het is dat het zo is. En als we dus hier. Laatste opmerking zijn willen dat er goede zorg goed werkt. En inderdaad mijn achterban misschien verrast wil dat de ook. Dan hoort er wel bij van onze kant. Als politici dat ook eerlijk zeggen wat er kan en wat er niet kan en een systeem wat goedkoper is eerlijker , meer zorg levert meer banen , minder wachtlijsten. Het bestaat niet als bestond had wat gemaakt en als u het wel heeft dan niet met praatjes komen van het staat op een website. Ik moet zeggen. De enige die eerlijk was volgens mij was Marijnissen die zei begin deze maand bij Paul. Dat is nog geen plan en ik ben heel erg bang dat de plannen niet is dat u niet eens weet wat de gevolgen zijn van uw plan voor de premie van de gemiddelde Nederlander van de belastingen die de gemiddelde Nederlander betaalt van de zorg die die krijgen. Het is er niet zolang dat er niet is. Kunt u beter over andere onderwerpen.

0:00 / 0:00 0.0

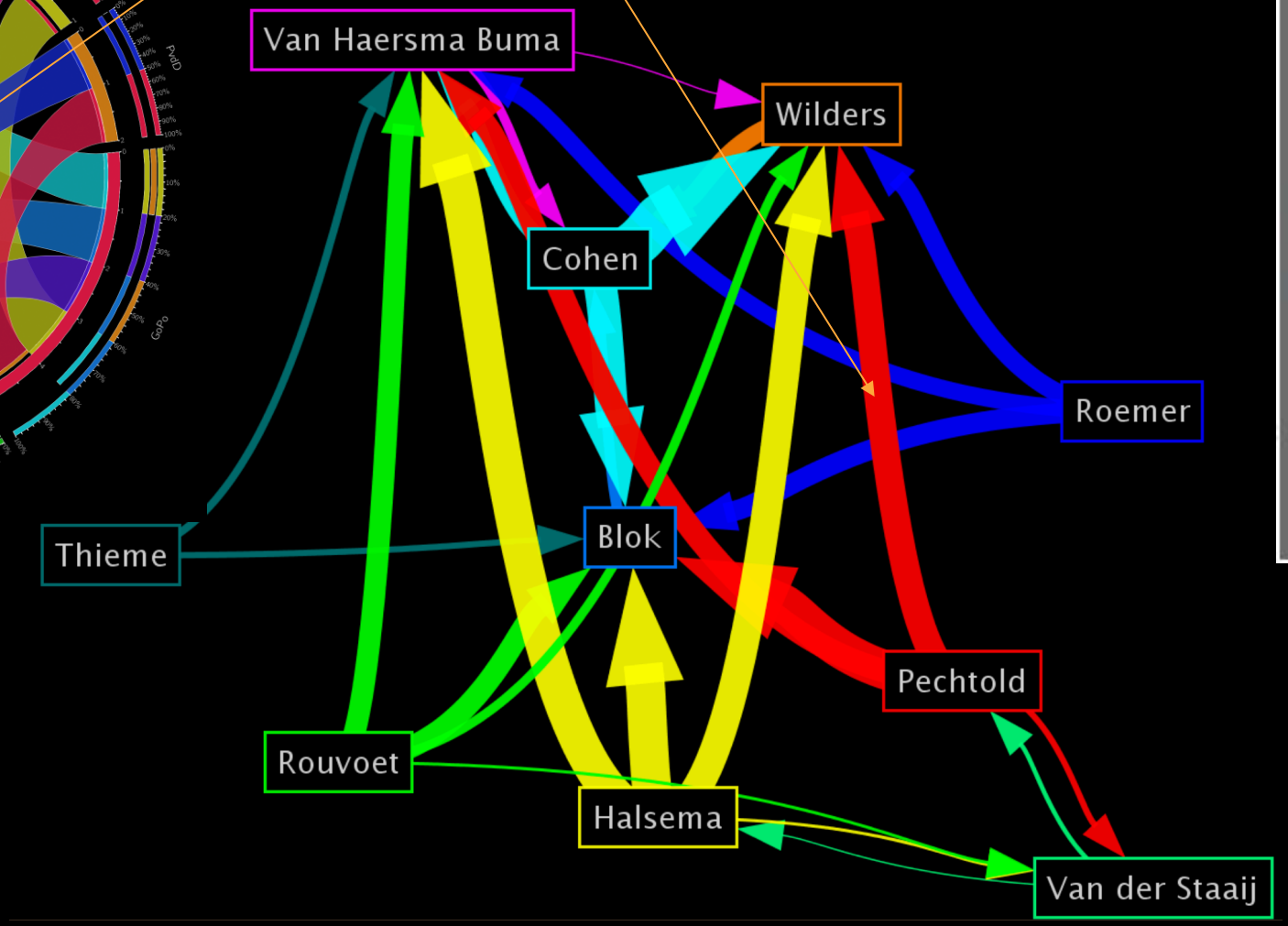
Tijdsindex Betrouwbaarheid Spreker

III - Interrupties per debat, gemiddeld

1. Mark Rutte	36,2
2. Jetta Klijnsma	21,9
3. Jet Bussemaker	21,1
...	
18. Jeanine Hennis	9,4
19. Lilianne Ploumen	9,0
20. Ivo opstelten	9,0

TELECATS

Click to select



TELECATS

Summary

Build an optimal ASR-engine for recognizing the Dutch Parliamentary speech

Figure out what can be done with the various by-product of the project

Train Moses to “translate” the spoken speech into written speech

Provide the reporters with an optimal editor for correcting the “translated” text

Figure out if it is (perceived as) useful

Questions?