Text-to-Pictograph Translation and Vice Versa

For People with Intellectual Disabilities

Leen Sevens
Vincent Vandeghinste
Ineke Schuurman
Frank Van Eynde

February 24, 2017

Centre for Computational Linguistics
KU Leuven

Content

- Introduction
- Text2Picto translation
- Picto2Text translation
- Able to Include
- Conclusion

Introduction

Augmentative and Alternative Communication

- Assist people with **Intellectual Disabilities**
- Allow them to **communicate** or **use the Internet**
- Increase life quality by **reducing social isolation**
- **Picture-based communication systems** are a form of AAC technology

Picture-based communication

Three types:

- Universal pictographs
- Emoticons
- Pictographs for people with disabilities

Picture-based communication

- Between **two and five million people** in EU
- Need for **picture-based communication interfaces** that enable social contact for illiterate and pre-literature users
- Interfaces should be
  - Easy to use
  - Configurable
  - Flexible

WAI-NOT Platform

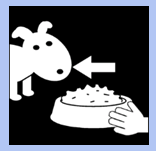
- Enabling internet access for people with mental disabilities
- Specific **applications** for these users
  - Chat and e-mail client with the Text2Picto engine
  - Easy-to-Read News
  - Text-to-Speech
  - Games
- www.wai-not.be

Pictograph sets

Sclera pictographs

hond (dog)

ik (I)



hond-eten-geven
(dog-food-feed)

afwas-proper-
terugbrengen
(dishes-clean-return)

Beta pictographs

hond (dog)

ik (I)

fruit-persen
(squeeze-fruit)

gebraden-kip
(roasted-chicken)

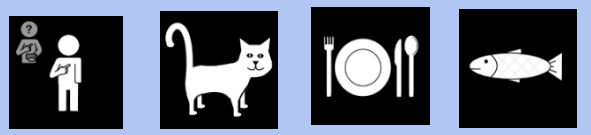
Text2Picto

Example message

Text

Mijn kat eet vis.
(My cat eats fish.)

Sequence of pictographs



Sclera

Beta

Original system (baseline)

- WAI-NOT's original system
- **Without any language technology**
- If input word matches **file name** of pictograph, a pictograph is shown
- Due to homonymy, this can be wrong!
- **Low coverage**
  - No morphological variation
  - Coverage of 41.33% for Beta (possibly wrong translations)
  - Much worse for Sclera: unusable
- BLEU score of 0.00% for Sclera
- BLEU score of 5.93% for Beta

WAI-NOT email corpus

- About 70.000 email messages sent with WAI-NOT
- **Three types of messages**
  - Emails written be **literate people** (teachers, parents,…)
    - Broad vocabulary
    - Hardest to translate
  - Short messages by **the intended users** (largest part)
    - No punctuation
    - No capitalization
    - Several spelling errors
  - **Noisy messages**
    - Random clicking on pictographs

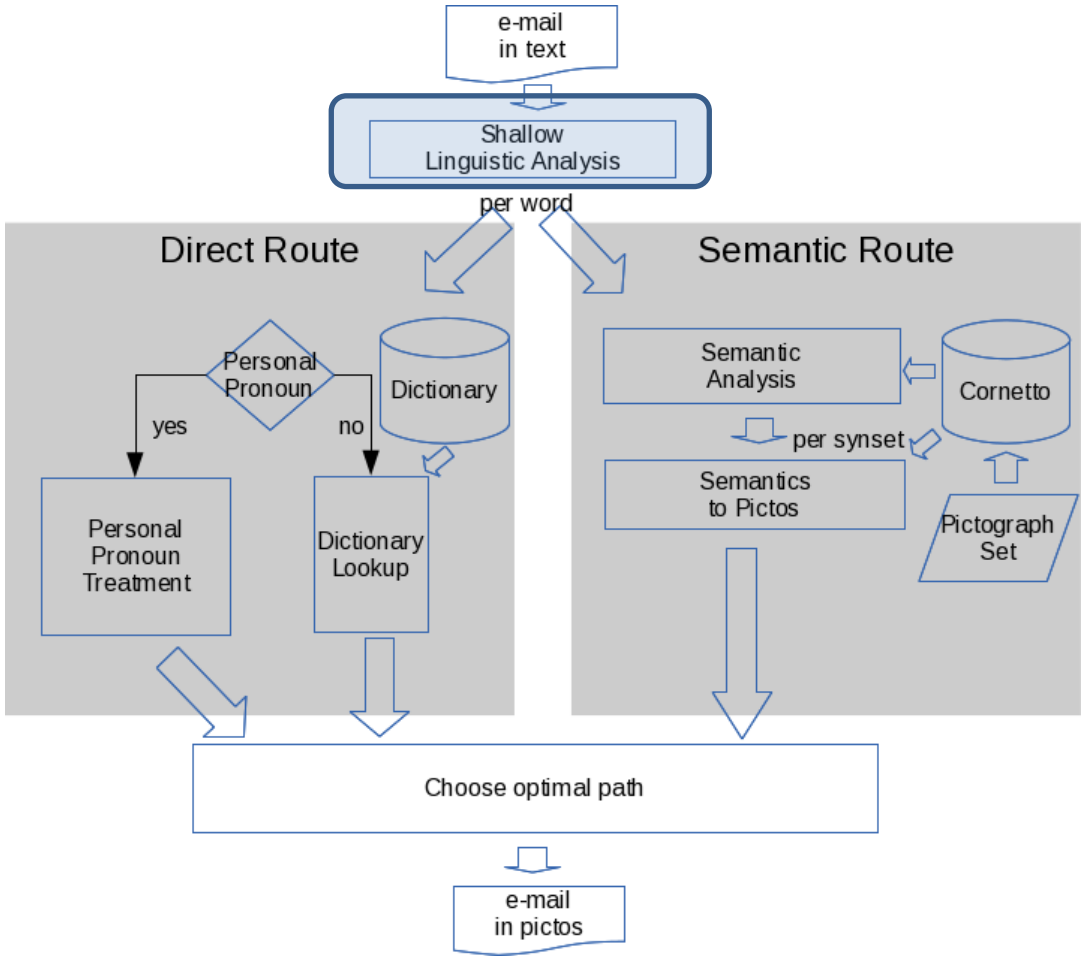
WAI-NOT email corpus

- **Development set**
  - 186 WAI-NOT messages
  - Used to test the system
  - Used to tune the system parameters
- **Evaluation set**
  - 50 messages
  - Average length: 20 words

Running example

“Hij is genezen” (He has recovered)

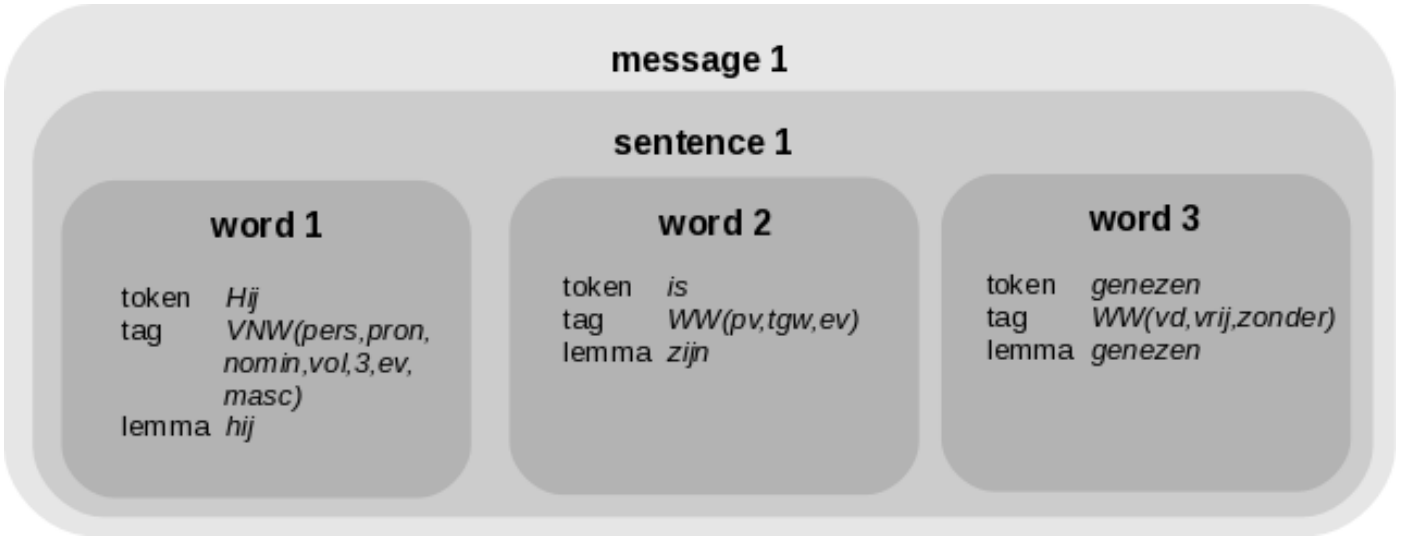
System description



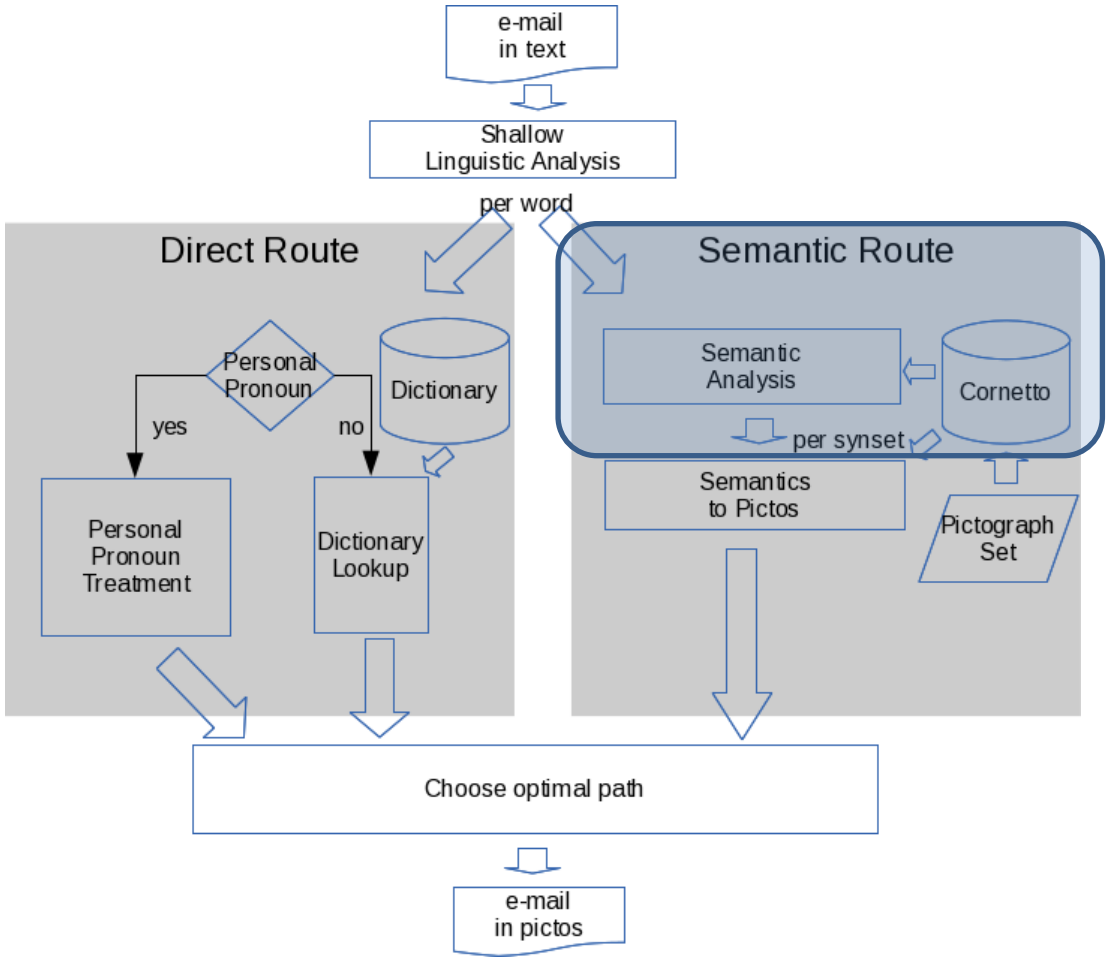
Shallow linguistic analysis

1. **Tokenization**
   - Splitting off all punctuation apart from hyphen/dash
2. **Basic spelling correction**
   - One deletion, one insertion, one substitution
3. **Part-of-Speech tagging**
   - HunPos tagger trained on manually corrected data (2 million words)
4. **Sentence detection**
   - Punctuation-based
5. **Separable verb detection (Dutch-specific)**
   - Verb + particle are sometimes written as separate words
6. **Lemmatization**
   - Look up token/tag combination in corpus
   - Else, apply regular expression rules

Example after shallow linguistic analysis

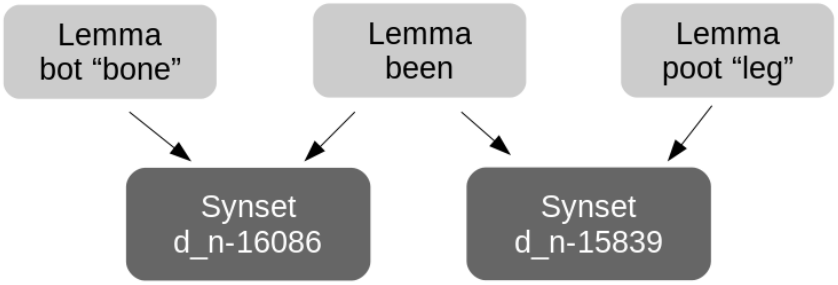


System description

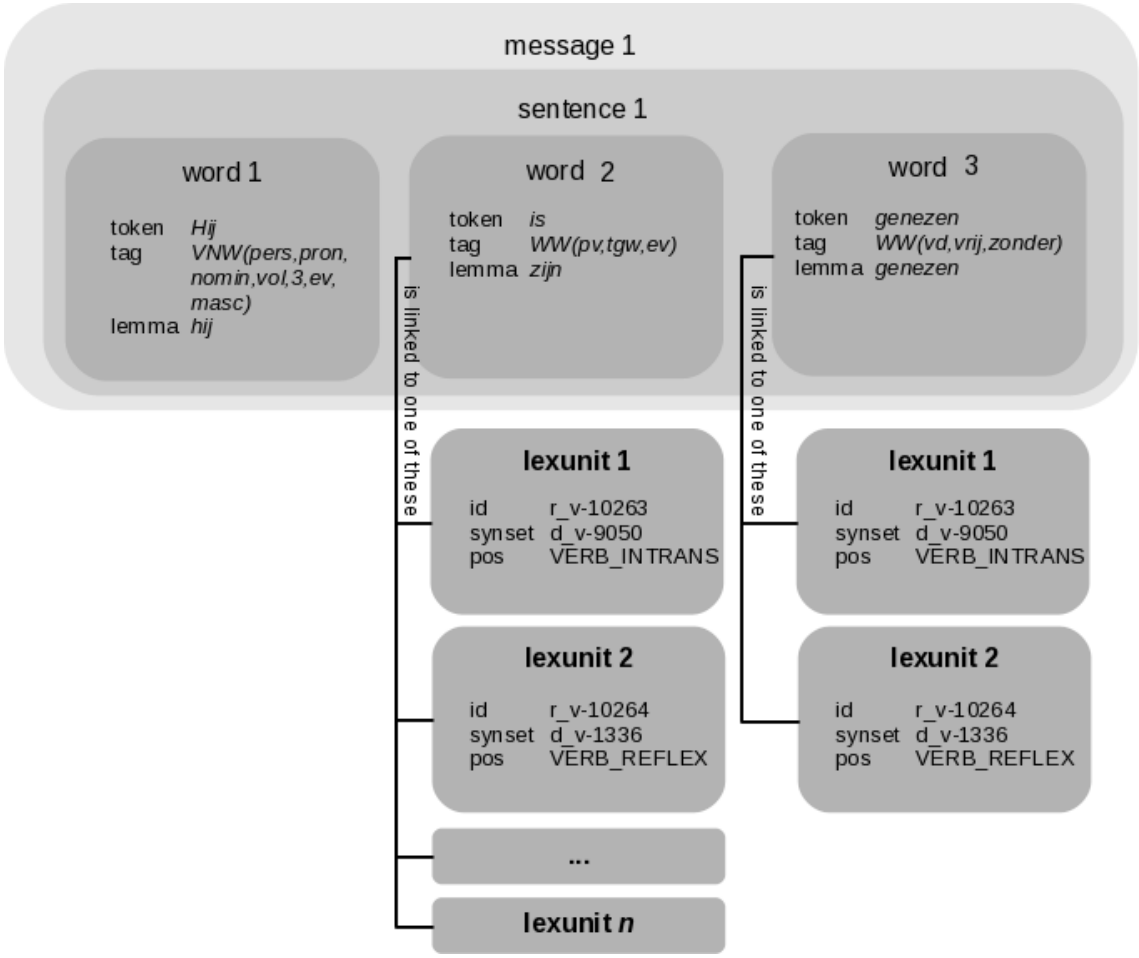


Semantic analysis

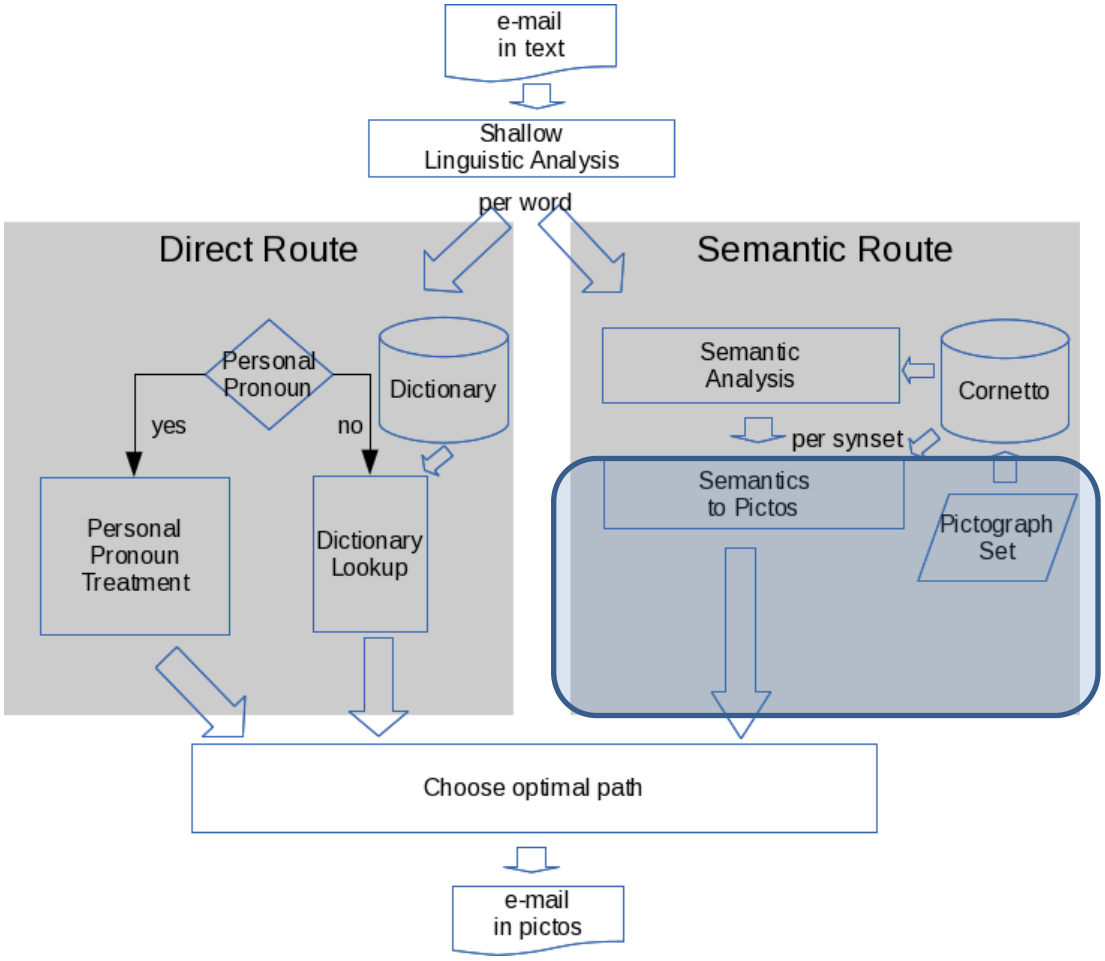
- Detect words indicating **negative polarity**
- Look up Cornetto **synsets** for each word
  - Like WordNet for Dutch
  - Contains relations between synsets (groups of synonymous words)
  - Synsets are linked to several **lemmas**
  - We added some Flemish lemmas



Example after semantic analysis

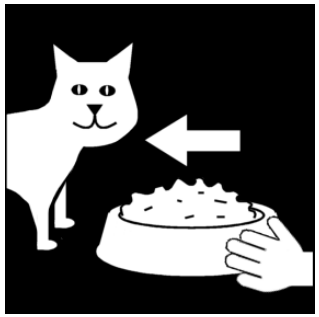


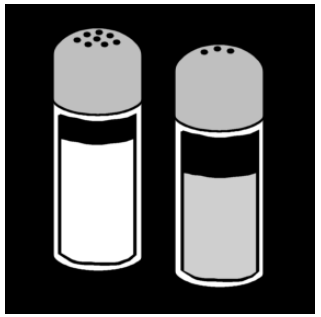
System description



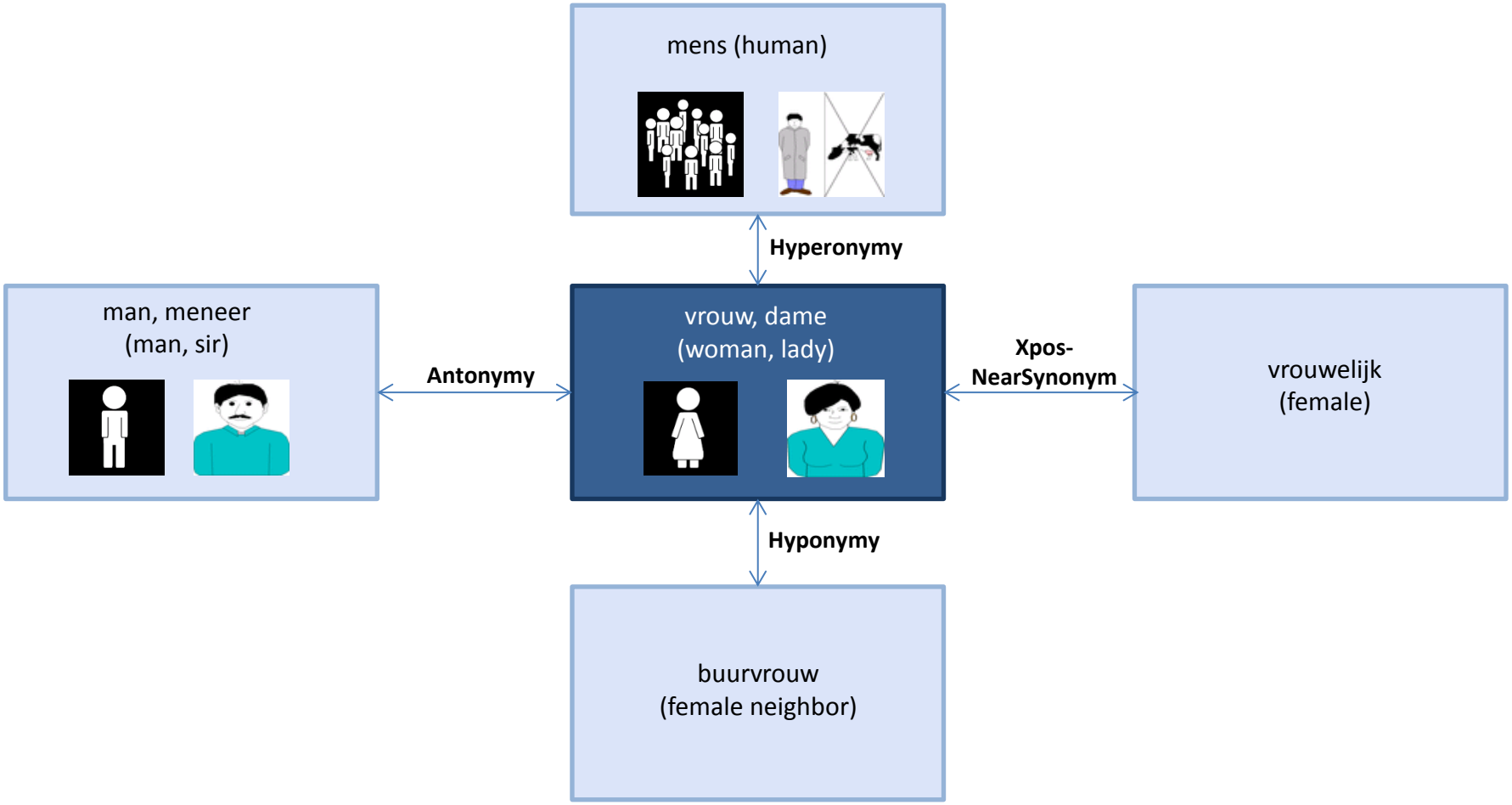
Linking pictographs to synsets

- **Manually linked** to Cornetto synsets
  - 5710 Sclera pictographs
  - 2760 Beta pictographs
- Sclera pictographs often depict **complex concepts**, such as
  - Verb + object
  - Noun + noun
  - …

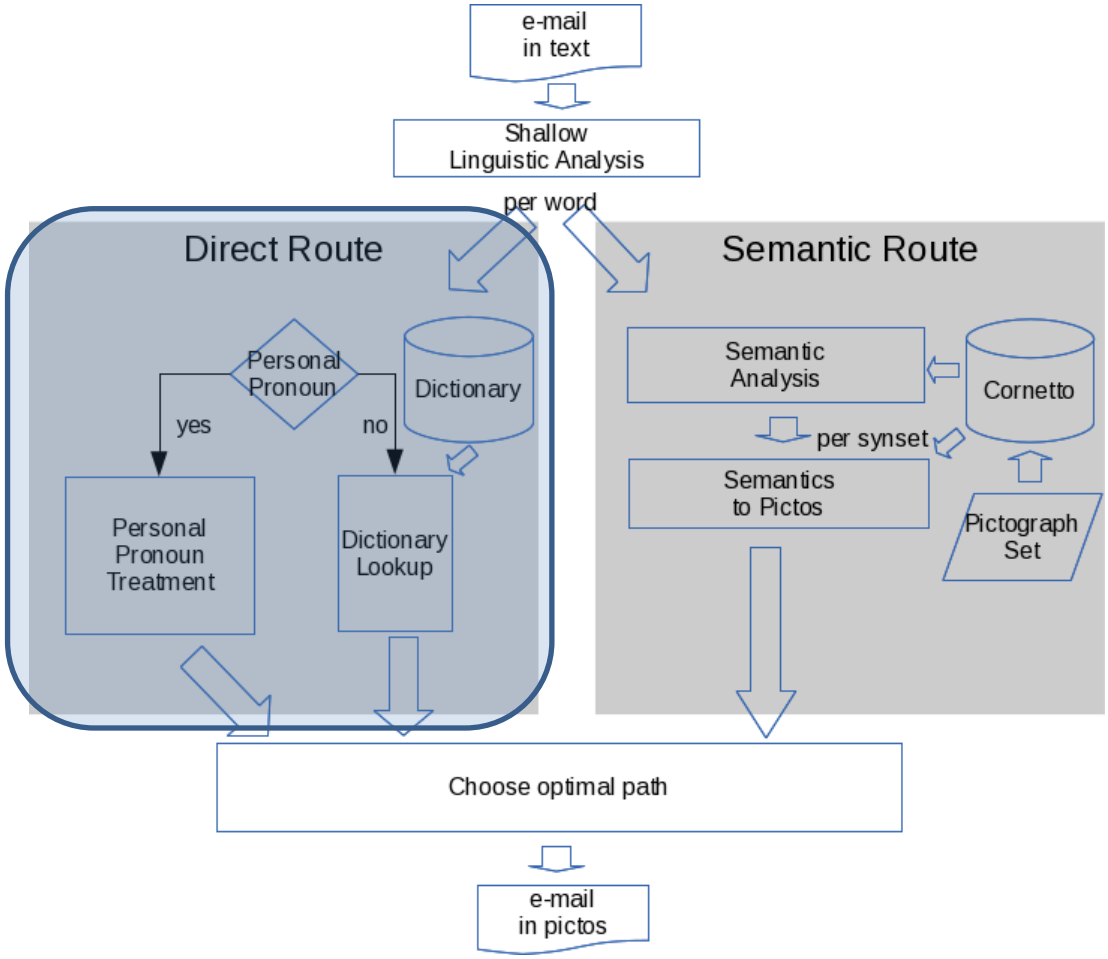




Using the synset links



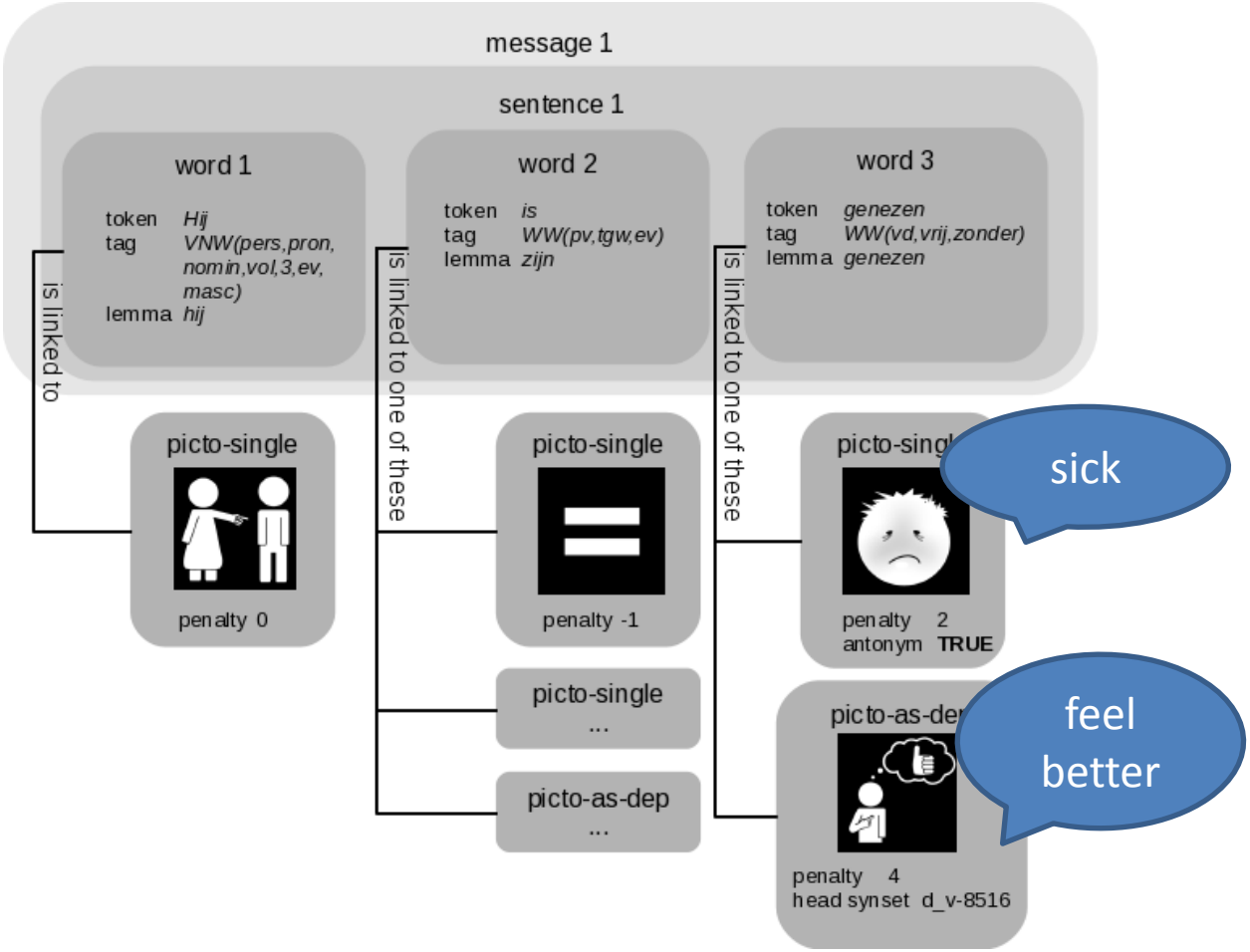
System description



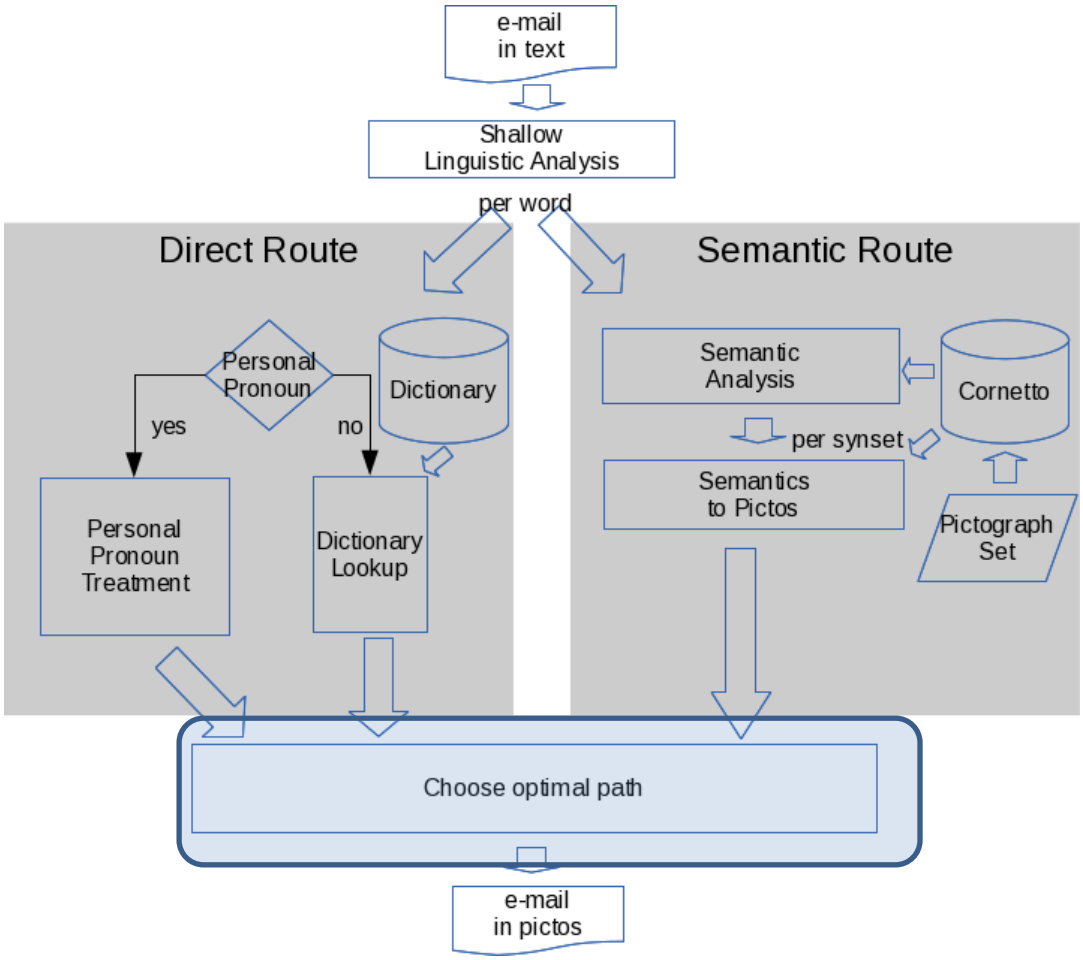
The direct route

- Not all words can be analyzed with Cornetto (only nouns, verbs, adjectives and adverbs)
- **Personal pronouns** are very frequent in e-mail messages and are not included in Cornetto: explicit treatment
- A translation mechanism that uses a **dictionary**
    - Token / tag / lemma -> instant translation into pictograph

Example after pictograph linking



System description



Finding the optimal path

- **A* search algorithm**
- Uses **parameters** (tuned beforehand using a local hill climber)
  - Maximum penalty threshold
  - Hyperonym penalty
  - XPos synonymy penalty
  - Antonymy penalty
  - Wrong number parameter
  - No number
  - Out-of-vocabulary parameter
  - Direct route advantage

Evaluation

- **50 messages** (83 sentences)
- We made **reference translations** in Beta and Sclera
- Translation with focus on how the content can best be expressed in pictographs

Automatic evaluation

- Progressively activating features of the system
- Evaluated using
  - **BLEU:** most used Machine Translation metric
  - **NIST:** similar to BLEU, but less credit to high-frequency non-informative n-grams
  - **WER:** word error rate
  - **PER:** position-independent word error rate
- Without/with (automatic) spelling correction

Automatic evaluation results

	No spelling correction				Automated spelling correction				Manual spelling correction			
Condition	BLEU	NIST	WER	PER	BLEU	NIST	WER	PER	BLEU	NIST	WER	PER
Sclera												
Baseline	00.00	1.43	96.27	92.13	00.00	1.38	99.00	95.58	00.00	1.84	97.51	94.20
Lemmatis.	01.87$^{\bullet}$	1.68†	94.48	89.36	01.91$^{\bullet}$	1.73†	93.92	88.81	02.44$^{\bullet}$	2.29†	92.27	87.02
Direct	10.74†	2.93†	75.41	69.20	11.57†	3.05†	74.59	68.09	14.17†	3.68†	71.96	65.88
Synonyms	**12.02**$^{\bullet}$	**3.32**†	**70.58**	**63.26**	**13.24**$^{\bullet}$	3.41†	**70.03**	**62.43**	**16.55**†	*3.97*†	*67.54*	*60.50*
Relations	11.44	3.29	72.24	64.50	12.75	**3.42**	71.41	63.26	16.12	3.96	68.78	61.33
Beta												
Baseline	05.93	2.29	80.76	72.21	04.94	2.40	81.10	71.99	04.70	2.81	79.19	69.85
Lemmatis.	07.77	2.90†	77.05	66.93	08.15	3.01†	76.94	66.14	10.14†	3.53†	74.92	63.78
Direct	11.96†	3.65†	66.59	57.48	12.72†	3.76†	66.14	56.47	16.98†	4.43†	63.44	53.77
Synonyms	16.57†	4.12†	**56.24**	**46.91**	18.70†	4.28†	**55.12**	**46.01**	23.01†	5.00†	52.42	43.31
Relations	**18.56**$^{\bullet}$	**4.22**†	56.47	47.24	**20.11**$^{\bullet}$	**4.40**†	55.46	**46.01**	**25.91**†	**5.17**†	**51.29**	**42.07**

$^{\bullet}p < 0.05$,$^{\dagger}p < 0.01$

Manual evaluation

- **One judge**
  - Remove untranslated non-content words to allow calculating **recall**
  - Judge for every translated word whether the pictograph is correct, to calculate **precision**

Manual evaluation results

Condition	Precision	With proper names		Without proper names	
		Recall	F-Score	Recall	F-Score
Sclera					
Baseline	77.60%	41.42%	54.01%	36.39%	49.55%
Text2Picto	89.24%	86.23%	87.71%	85.18%	87.16%
Rel. improv.	*15.00%*	*108.19%*	*62.39%*	*134.06%*	*75.92%*
Beta					
Baseline	82.73%	62.23%	71.03%	59.57%	69.27%
Text2Picto	85.91%	89.45%	87.64%	88.68%	87.27%
Rel.improv.	*3.84%*	*43.73%*	*23.38%*	*48.88%*	*26.00%*

Improvement #1: syntactic simplification

Ik ben woensdag een lieve baby gaan bezoeken en ik heb hem een papfles gegeven.
I am Wednesday a cute baby go visit and I have him a bottle given.
'On Wednesday I went to visit a very cute baby and I gave him the bottle.'

Improvement #1: syntactic simplification

Ik ben woensda ɔapfles gegeven.
I am Wednesd ɔottle given.
'On Wednesday e.'

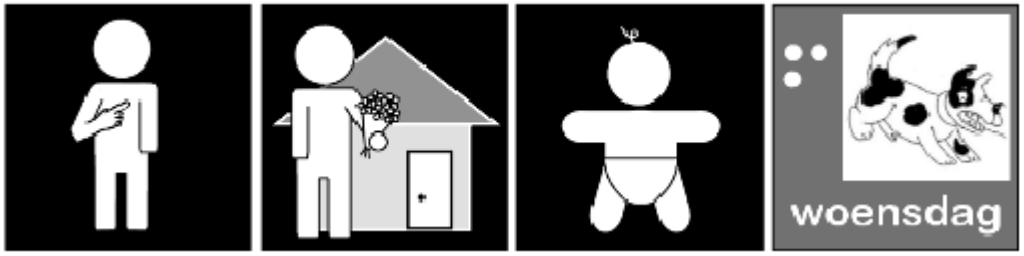
Improvement #1: syntactic simplification

- **Solution:** automated syntactic simplification for pictograph translation:
  - Split long input sentences into several shorter ones
  - Convert passive constructions into active constructions
  - Adhere to Subject-Verb-Object (SVO) order
  - Convert relative clauses into independent clauses
  - Etc.

Improvement #1: syntactic simplification

Ik ben woensdag een lieve baby gaan bezoeken en ik heb hem een papfles gegeven.
I am Wednesday a cute baby go visit and I have him a bottle given.
'On Wednesday I went to visit a very cute baby and I gave him the bottle.'

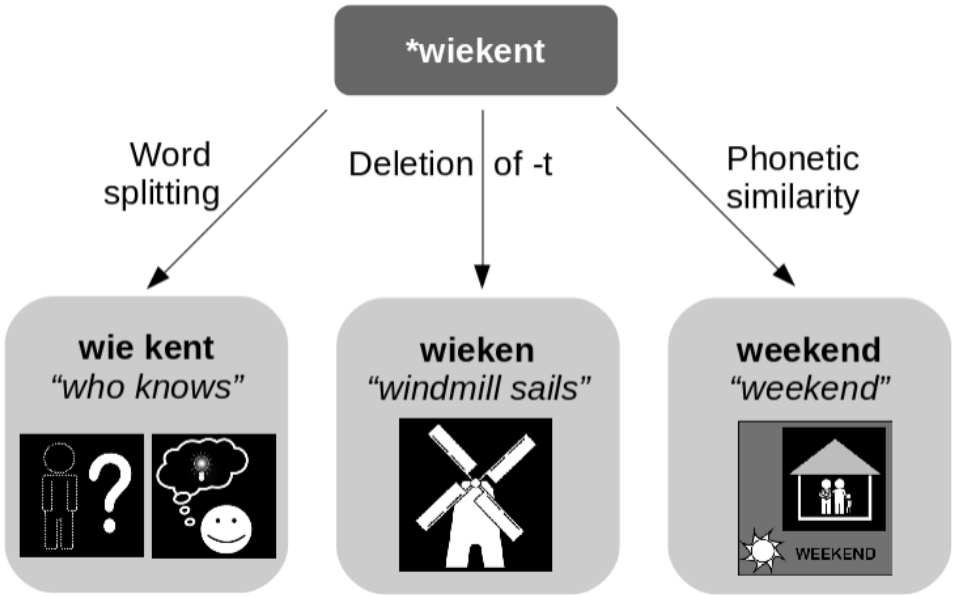
(a)



(b)

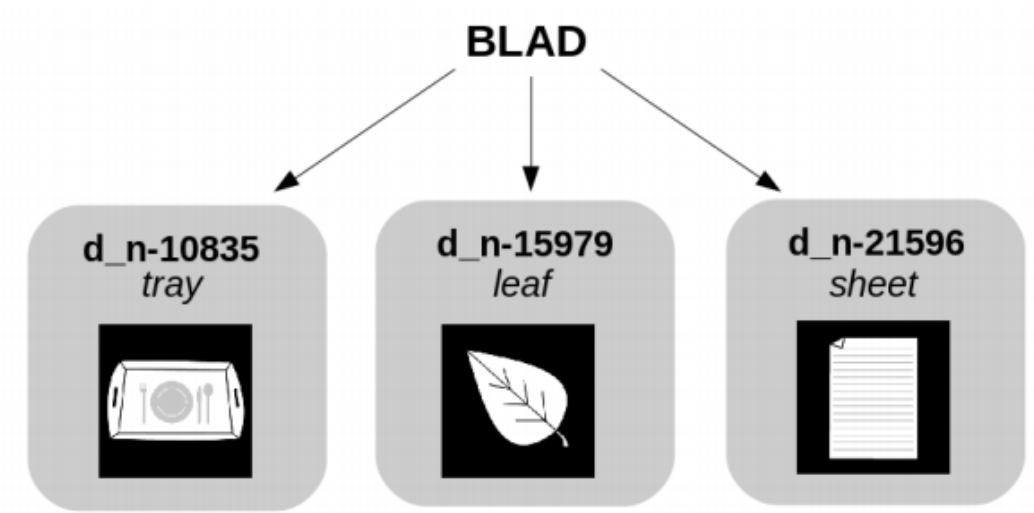
Improvement #2: improved spelling correction

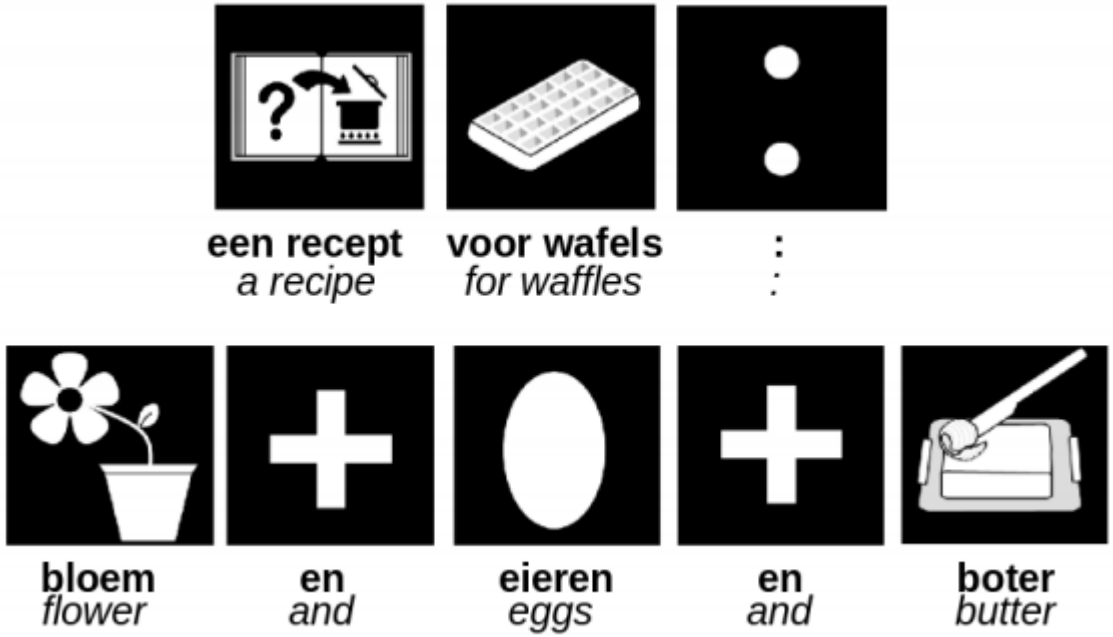
- **New method:**
  - Generate spelling variants (for both phonetic errors and typographic errors)
  - Fuzzy matching (Machine Translation) techniques for finding the best combination of spelling variants



Improvement #3: word sense disambiguation

- The original system did not yet select the appropriate sense of a word
- The most frequent sense was chosen, which resulted in incorrect translations



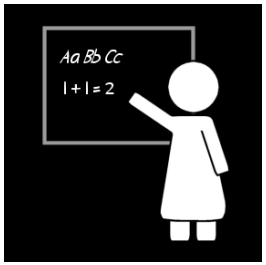

een recept
a recipe
voor wafels
for waffles
:
:
bloem
flower
en
and
eieren
eggs
en
and
boter
butter

Delicious. ;-)

Improvement #3: word sense disambiguation

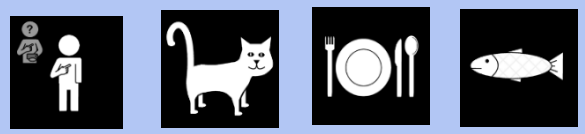
- We implemented a **word sense disambiguation (WSD)** tool (created by Vossen et al. (2010) within the DutchSemCor project)
- The WSD scores are now added as new features of the synsets in the Text2Picto engine
- A high WSD score biases the selection of the pictograph toward the winning sense

Picto2Text



Example message

Sequence of pictographs



Sclera

Beta

Text

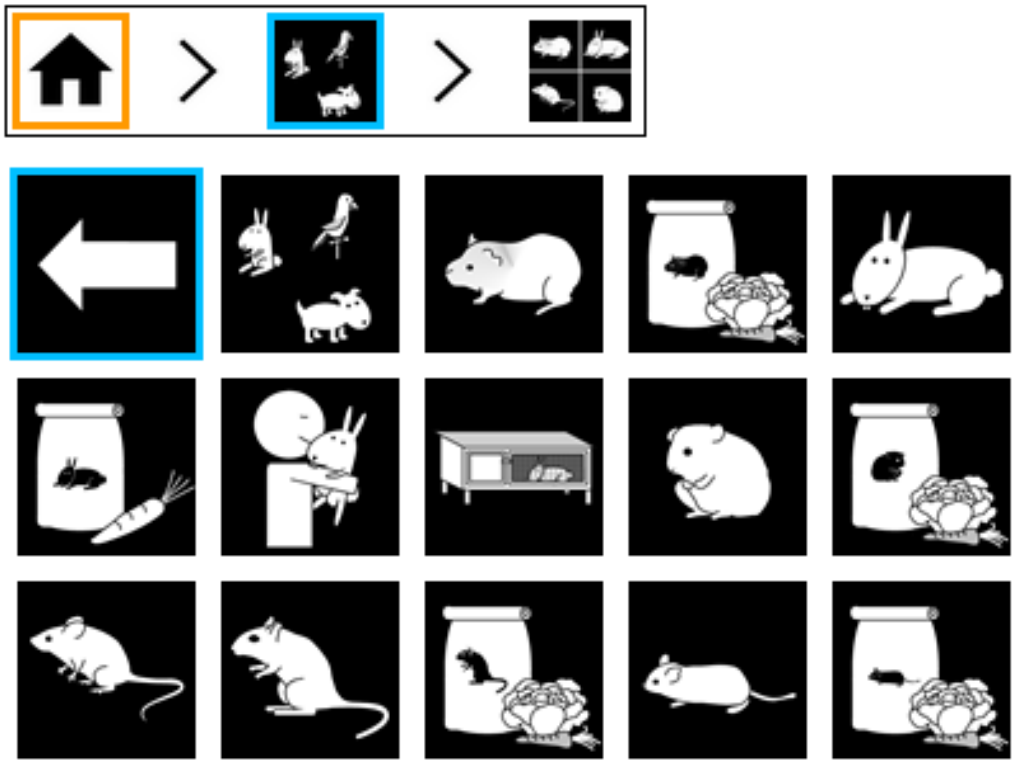
Mijn kat eet vis.
(My cat eats fish.)

Input methods

- **Original version (WAI-NOT):**
  - Two-level static hierarchy
  - Too many unordered pictographs on lowest level
  - Categorial inconsistency
- **New approach:**
  - Three-level static hierarchy
  - Dynamic pictograph prediction

Three-level static hierarchy

- **Top-level categories:**
  - Topic detection applied on WAI-NOT corpus
  - Latent Dirichlet Allocation
- **Subcategories:**
  - Based on WordNet (Cornetto) relations
- **Lowest level:**
  - Ordered by frequency in WAI-NOT corpus
  - Overruled by logical ordering: numbers, months, days, pairs of antonyms



Dynamic pictograph prediction

- **N-gram based prediction**
  - WAI-NOT corpus was automatically translated into Beta / Sclera pictographs (280K pictographs)
  - Build language models using SRILM for Beta / Sclera
  - Present the most likely next pictograph based on the two previous pictographs
- **Word association based prediction**
  - Retrieve a list of semantically similar words with the DISCO tool and translate them into pictographs

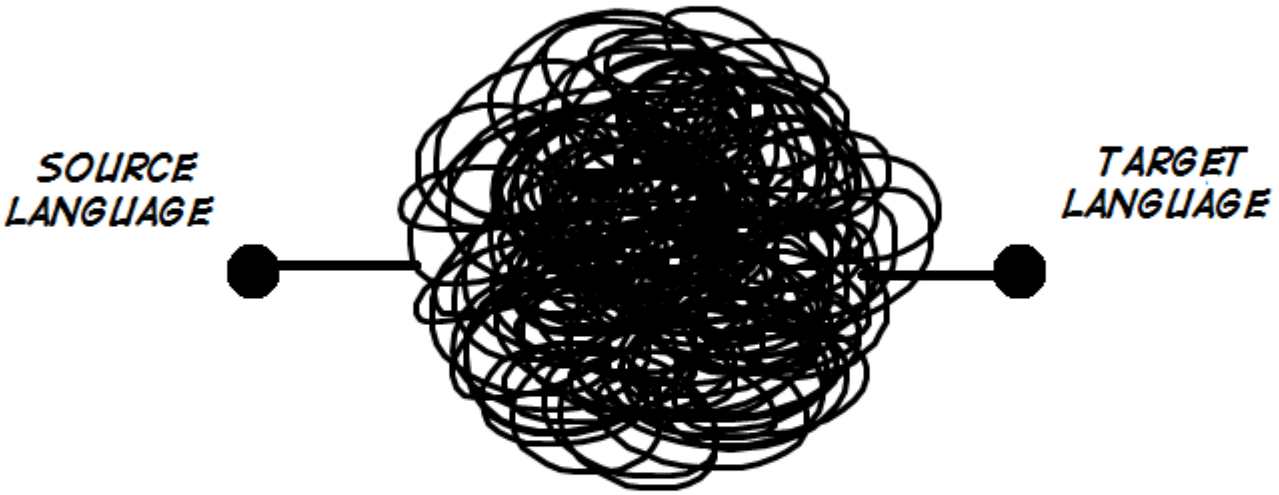
The user inserts pictographs

The prediction tool suggests new pictographs

to bark, to eat, to be,...

Pictograph to Natural Language Generation

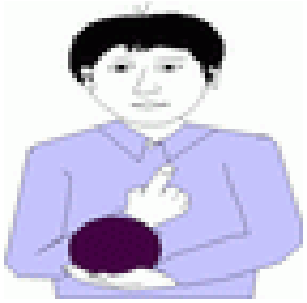
- **Difficulties**
  - Pictograph-for-word correspondence will almost never provide acceptable output
  - Pictograph languages often lack pictographs for function words
  - A single pictograph often encodes information corresponding to multiple words with multiple inflected word forms



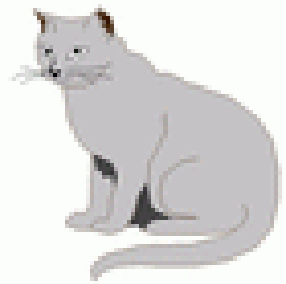
THE PROCESS OF MACHINE TRANSLATION.

System architecture

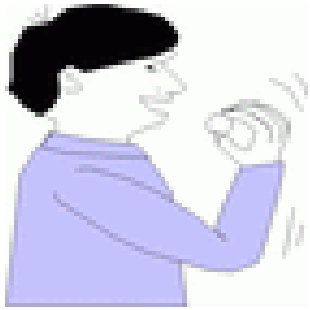
- **Step 1.** Take the file names of the pictographs



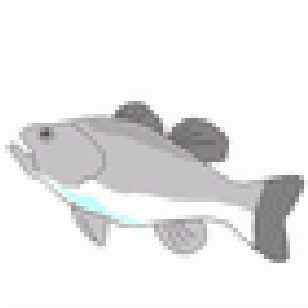
mijn.png



kat.png

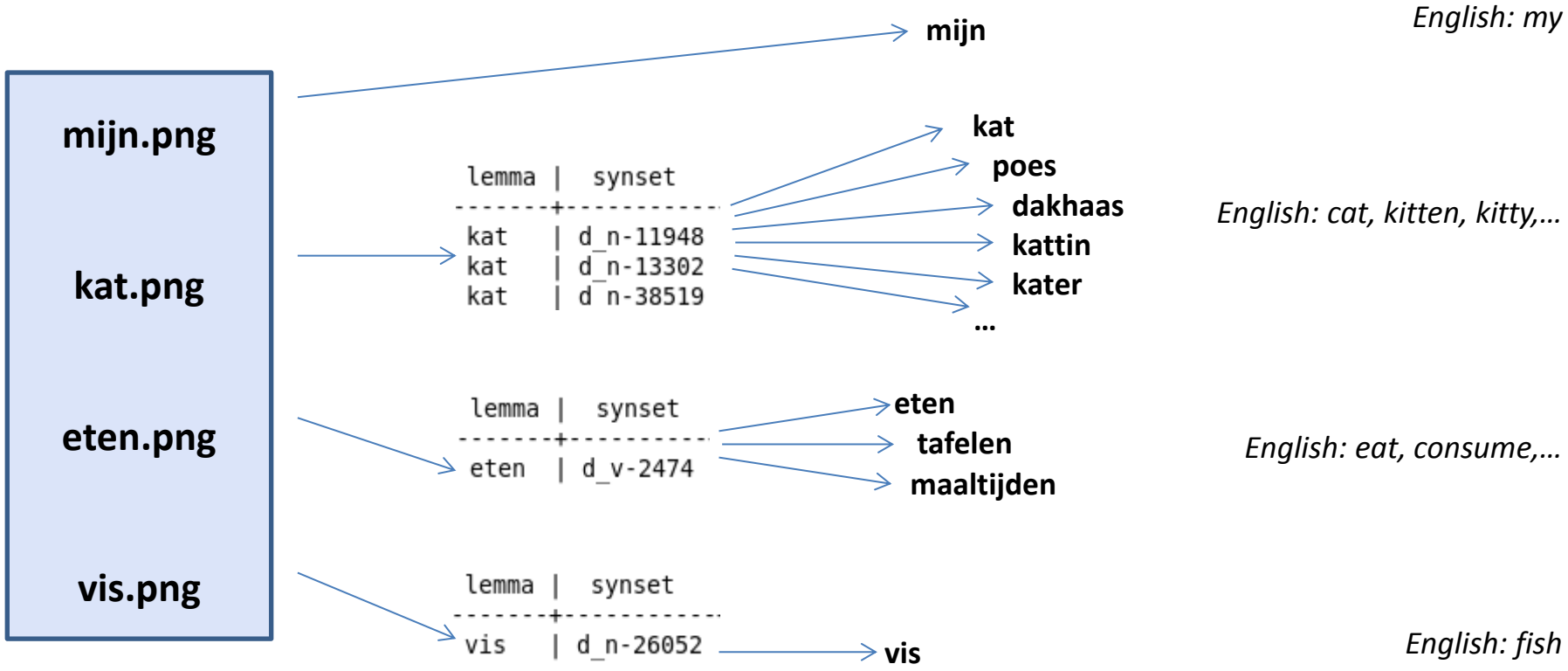


eten.png

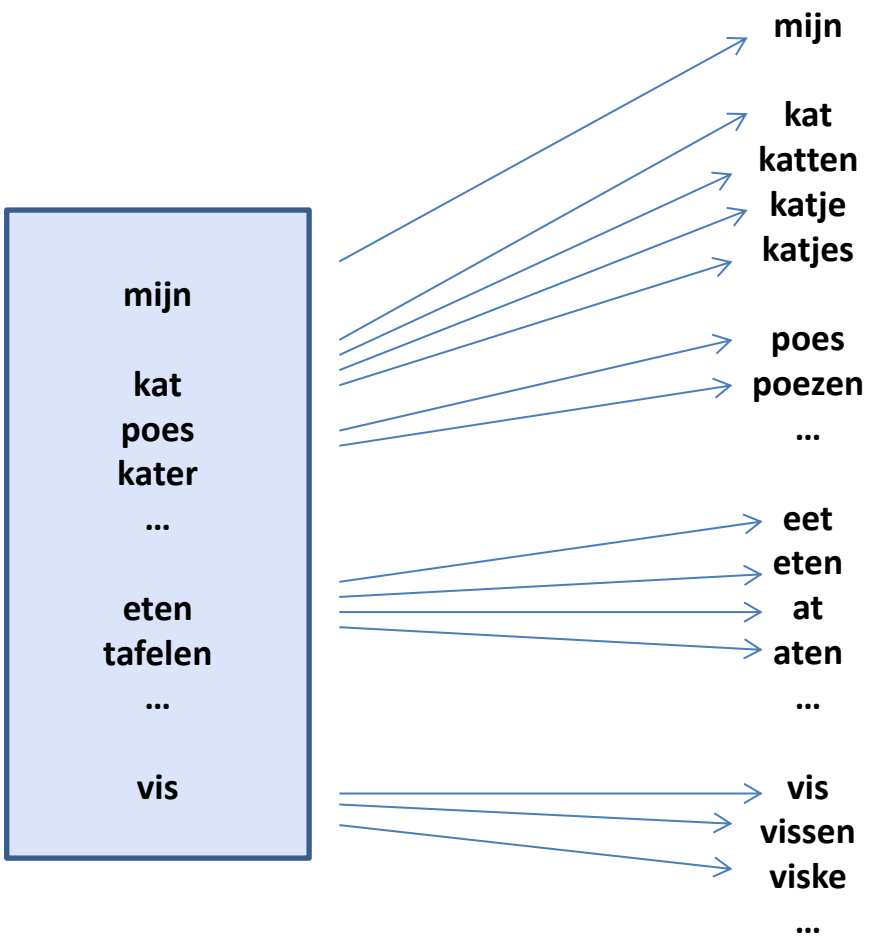


vis.png

- **Step 2.** Find the synsets that are connected to these file names and retrieve all the lemma's that are contained within these synsets



- **Step 3.** For every lemma, generate its paradigm (reverse lemmatization)



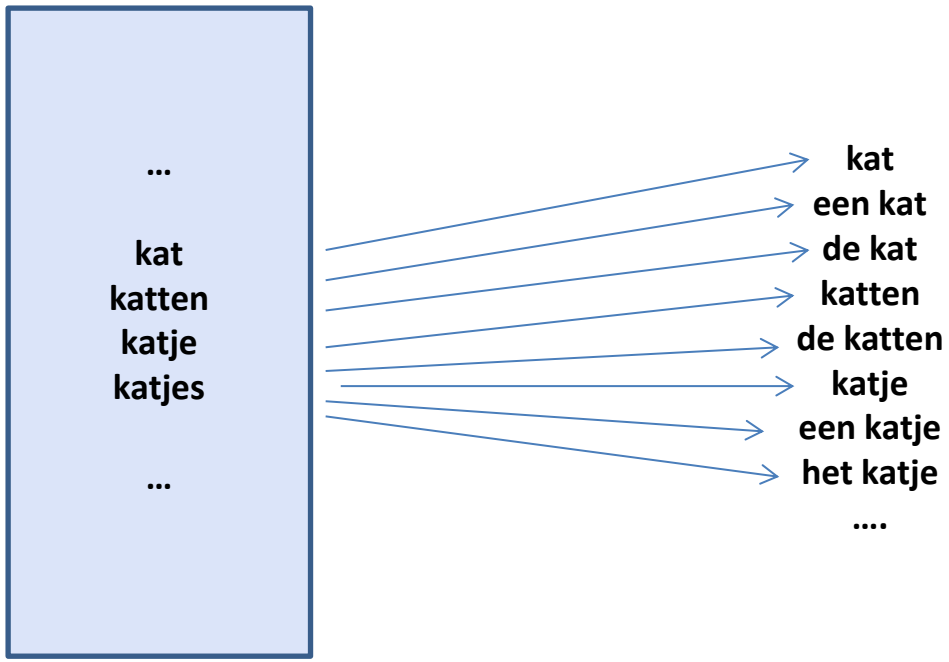
English: my

English: cat, cats, kitten, kittens,…

English: eat, eats, ate, consumed,…

English: fish

- **Step 4.** For every noun, create variants with/without article



English: the cat, a cat, cat, the cats,…

- **Step 5.** Find the **most likely combination** of all these words by using an A* algorithm, based on a **trigram language model** (trained on very large Dutch corpora with the SRILM toolkit)
  - Europarl v.3 corpus (Koehn, 2005), the CGN (Oostdijk and Broeder, 2002), the CLEF corpus (Peters and Braschler, 2001), the DGT corpus (Steinberger et al., 2012) and Wikipedia entries

Output:
Mijn kat eet vis.
(My cat eats fish.)

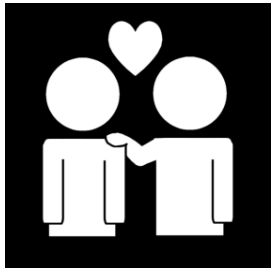
Results

Condition	BLEU	NIST	WER
Sclera			
Baseline	0.0175	1.5934	76.4535
Rev. lem.	0.0178	1.6852	76.8411
Direct	0.0420	2.2564	66.9574
Synsets	0.0535	2.5426	65.9884
Articles	0.0593	2.8001	67.4419
Beta			
Baseline	0.0518	2.767	70.4457
Rev. lem.	0.0653	3.0553	70.3488
Direct	0.0814	3.3365	63.0814
Synsets	0.0682	3.1417	61.4341
Articles	0.0739	3.4418	63.1783

Conclusions about Picto2text

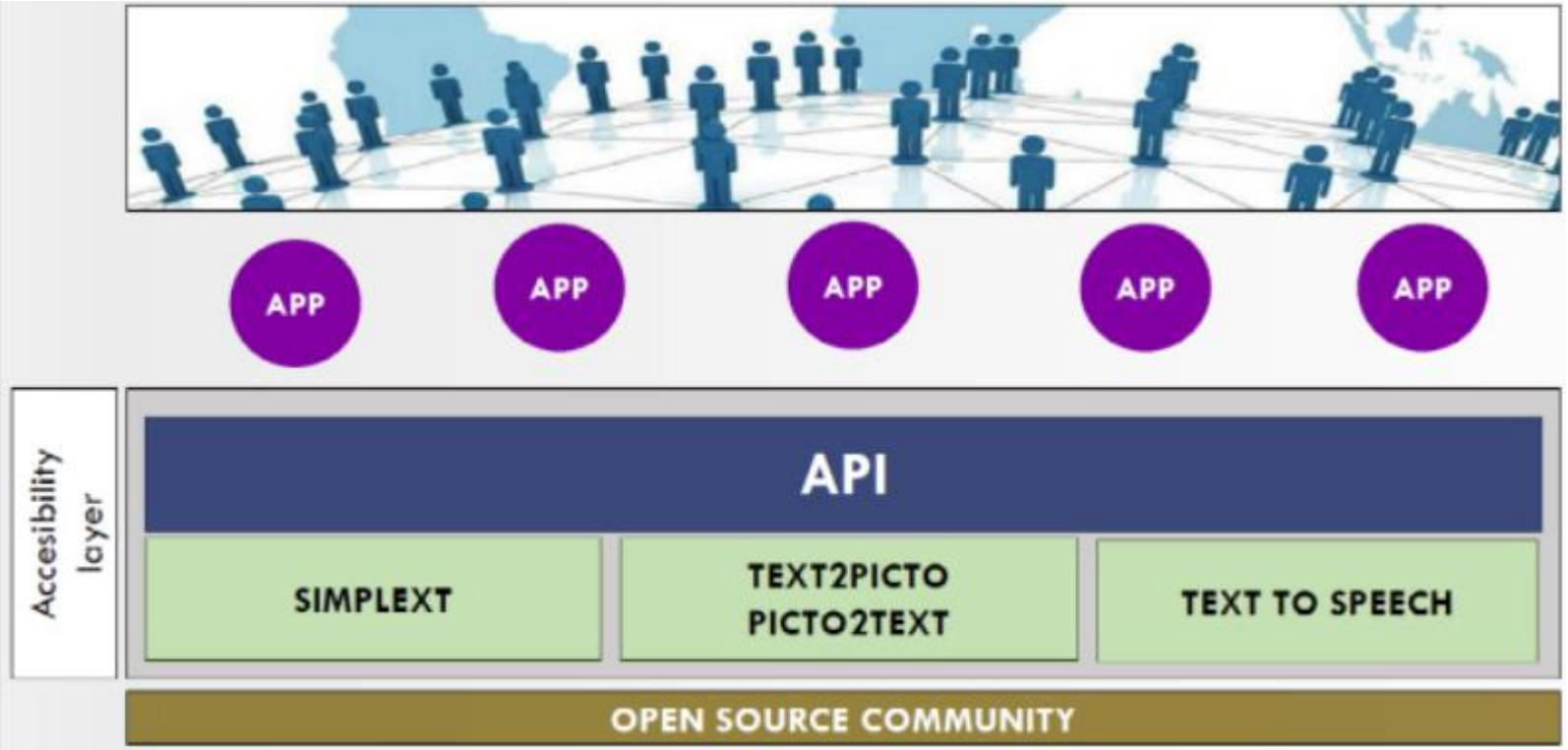
- Large improvement over baseline
- Ample room for further improvement
- **Future improvements** (to do):
  - Explore neural net-based models
  - Explore rule-based models
  - Explore hybrid models
  - Automated grammar correction

Able to Include



Project description

- European project: March 2014 – April 2017
- **Goal:** Improve the living conditions of people with Intellectual Disabilities (ID)
  - Natural Language Processing tools can bring independency to people with ID
  - Ex. Social media websites, email
- **Accessibility Layer** based on three key technologies:
  - **Tech 1:** Text and content simplifier (Simplext): *English & Spanish*
  - **Tech 2:** Text-to-speech functionalities: *Dutch, English & Spanish*
  - **Tech 3:** Text-to-Pictograph and Pictograph-to-Text translation: *Dutch, English & Spanish*



- The Accessibility Layer can be used in combination with a number of other applications, such as Facebook (on smartphones and tablets)
- It is **context-aware and customizable**

Extending the tool toward other languages

- **"Language-independent"** design of the Text2Picto and Picto2Text tools:
  - The systems can easily be extended toward other languages
  - We made **English and Spanish** versions of the tools (Russian and French: in progress)
  - Thanks to the use of **WordNets**: easy transfer of pictographs to words in other languages
  - Add language-specific linguistic resources
  - Deal with language-specific issues

User tests

- Our partners (Belgium, UK, Spain) are testing the tools in three **user scenarios**:
  - **Labor integration**: email client, word processor, pdf reader
  - **Mobility:** guidance systems for people with ID
  - **Leisure within information society:** Facebook, Twitter, WhatsApp,…

User tests

User tests

- The outcome of the pilots affects our approach:
  - Ex. The need for **decent pictograph input methods**
  - Ex. The users wanted to see the original input words written next to the pictographs in the Text2Picto tool: **PictoParallel**
  - Ex. Long pictograph translations turned out to be rather confusing for the users: **syntactic simplification methods**
  - Ex. We are constantly **adding new pictographs** (on request)

Conclusion

Up to now

- **Large improvements** over baseline
  - Tested in vitro
  - Partly tested in vivo
  - Make system usable in real life
- Expansion to **new languages**
  - If there is a WordNet
  - If there are basic linguistic resources available
  - Not too hard

Future and ongoing

- Continue in vivo testing
- Improve Picto2Text
- Think of new applications, new target groups (the elderly, immigrants,…)

Thank you!

Contact:
leen.sevens@kuleuven.be
http://ccl.kuleuven.be/~leen/

Picto demo & publications:
http://picto.ccl.kuleuven.be