FACULTY OF ARTS
AND PHILOSOPHY

DEPARTMENT OF TRANSLATION INTERPRETING AND COMMUNICATION
LT3 LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

SEMANTIC ANALYSIS OF TEXT: USE CASES (CYBERBULLYING DETECTION, IRONY DETECTION, ASPECT-BASED SENTIMENT ANALYSIS, COGNATE DETECTION, WINE CLASSIFICATION)

Els Lefever

UCLouvain, March 5th 2020

GHENT
UNIVERSITY

language and
translation
technology
team

LT³, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

GHENT UNIVERSITY

LT³

•Dpt of Translation, Interpreting and Communication, Faculty of Arts and Philosophy, Ghent University

•fundamental and applied research in **language and translation technology** > How can we build models for computational natural language understanding?

•3 ZAP, 4 Postdocs, 11 Phd students

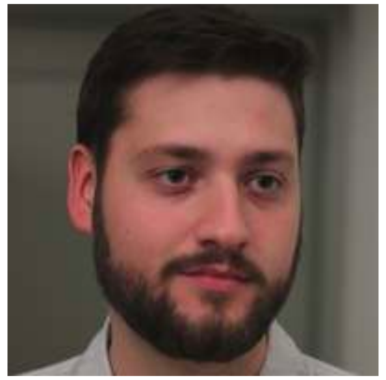
•Headed by Prof. Véronique Hoste

GHENT UNIVERSITY

TERMINOLOGY & COMPUTATIONAL SEMANTICS

- Lead: Prof. Els Lefever

- Automatic terminology extraction from monolingual, bilingual and comparable corpora (Ayla Rigouts Terryn)
- Term ambiguity in interdisciplinary research (Julie Mennes)
- Semantic Interoperability in medical communication between physicians and patients (Dirk Van Nimwegen)
- PLATOS: Detection of topics, stance and argumentation in a social media corpus (Nina Bauwelinck)
- SENTiVENT: event extraction and sentiment analysis for financial news (Gilles Jacobs)
- Automatic hypernym detection, automatic cognate detection, linguistic preprocessing (Els Lefever)



GHENT UNIVERSITY

WWW.LT3.UGENT.BE/TOOLS/

GHENT UNIVERSITY

language and translation technology team

NEWS PROJECTS PUBLICATIONS PEOPLE TOOLS TEACHING

Home > Tools

annotation

- → Sort by readability (NL)
- → Sort by readability (EN)
- → Expert Readers (NL)
- → Expert Readers (EN)

demo

- → Compound splitter
- → Readability
  - → Assessing Readability
  - → Classical formulas
  - → Machine learning
- → Normalisation Demo
- → Sentiment Demo
- → LeTs Demo

software

- → LeTs

CVT.UGENT.BE

UNIVERSITEIT GENT

Department of Translation, Interpreting and Communication

CENTRUM VOOR TERMINOLOGIE - GENTERM - TERMINOLOGY CENTRE

HOME TEACHING RESEARCH SERVICES *NEWS* GENTERM LINKS DOWNLOADS CONTACT

Projects

- MeSH Termbase
- EDiCT
- JuriGenT
- IATE-CvT

CvT

New tern

The **Terminology Centre** (CvT) is active within the Department of Translation, Interpreting and Communication of Ghent University. The CvT co-ordinates the Department's activities on terminology and terminography.

These activities relate to teaching, research and services.

The CvT's **staff** is drawn from the Department's language sections and the language technology section.

The CvT operates as a unit within the research group LT3 the Department's Language and Translation Technology Team. Whereas most other research within LT3 is related to tools development, the CvT focuses on the use of tools for terminology management, term recognition, term extraction etc. and on the manual compilation of termbases.

Core Members

- Els LEFEVER (LT3, Head)
- Joost BUYSSCHAERT (English) (hon. Head)
- Bart DEFRANCQ (French)

WWW.GHENTCDH.UGENT.BE

GHENT UNIVERSITY

GHENT CENTRE FOR DIGITAL HUMANITIES

HOME | SERVICES | PROJECTS | ACTIVITIES | ABOUT THE CENTRE

The GhentCDH

The Ghent Centre for Digital Humanities (GhentCDH) engages in the field of 'Digital Humanities' at Ghent University, ranging from archaeology and geography to linguistics and cultural studies. It develops DH collaboration and supports research projects, teaching activities and infrastructure projects across the faculties.

Geospatial analysis

The Ghent CDH offers advice, support and training regarding geospatial data management, analysis and visualisation to the humanities and social sciences researchers at the Ghent University.

Read more about this service

DH Support | Geospatial analysi

Outline

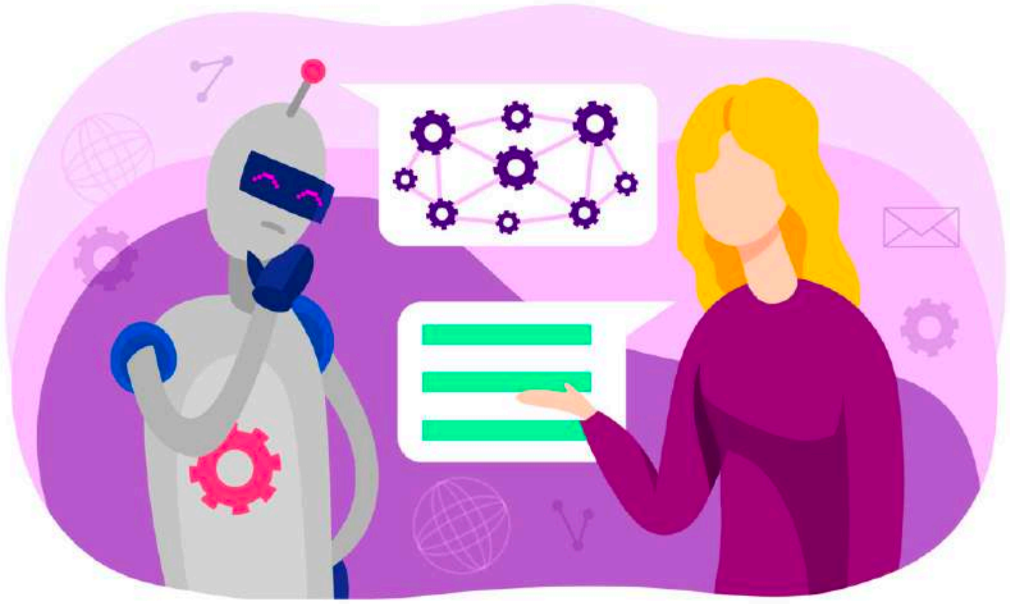
1. NLP & semantic analysis of text

2. Use cases:
- Cross-lingual Word Sense Disambiguation
- cyberbullying detection
- irony detection
- sentiment analysis
- cognate detection
- wine classification
- other …

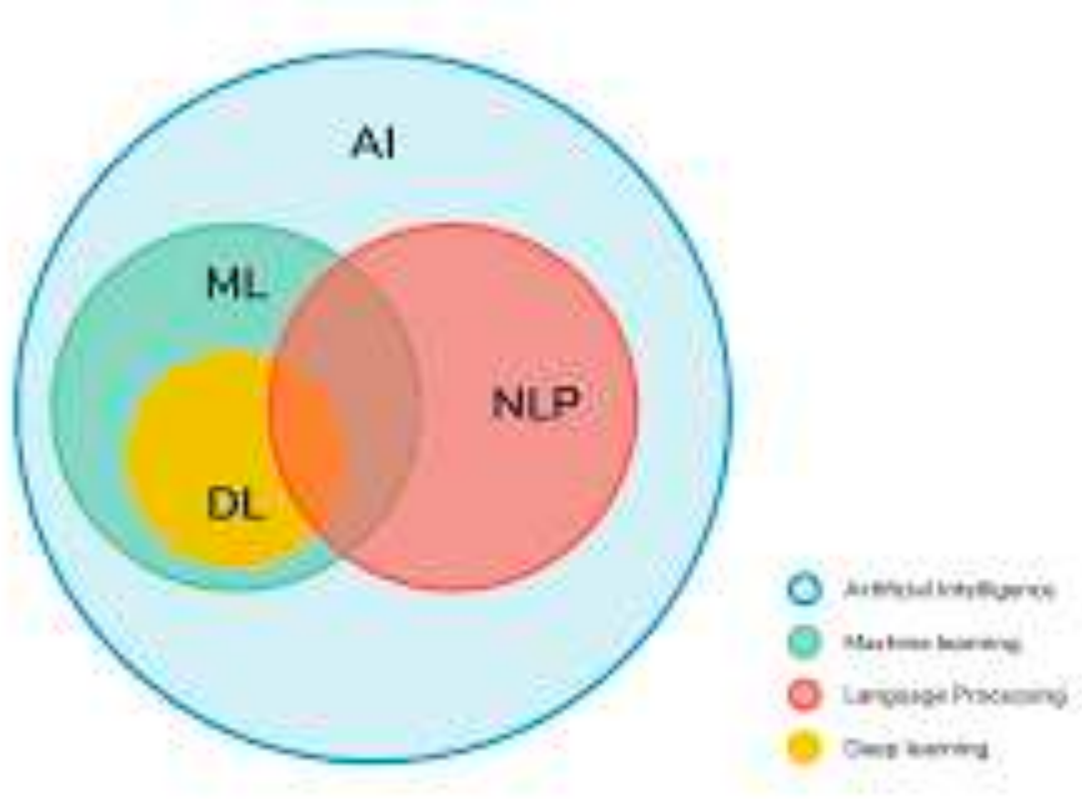
NLP & SEMANTIC ANALYSIS OF TEXT

NATURAL LANGUAGE PROCESSING (NLP)

Subfield of linguistics, computer science, … and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data (Wikipedia.org)



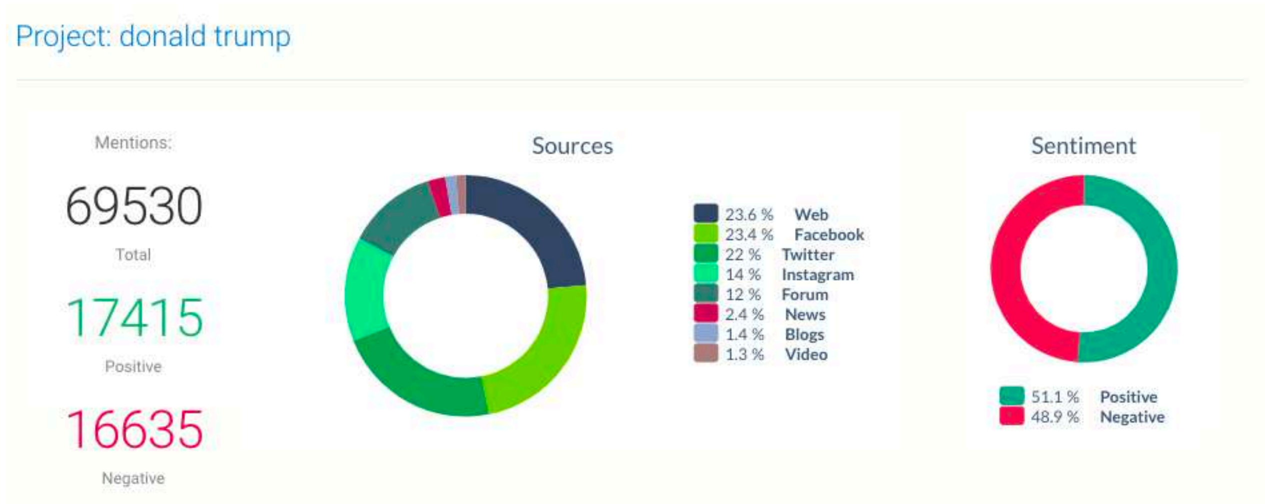
Ref: Landbot.io



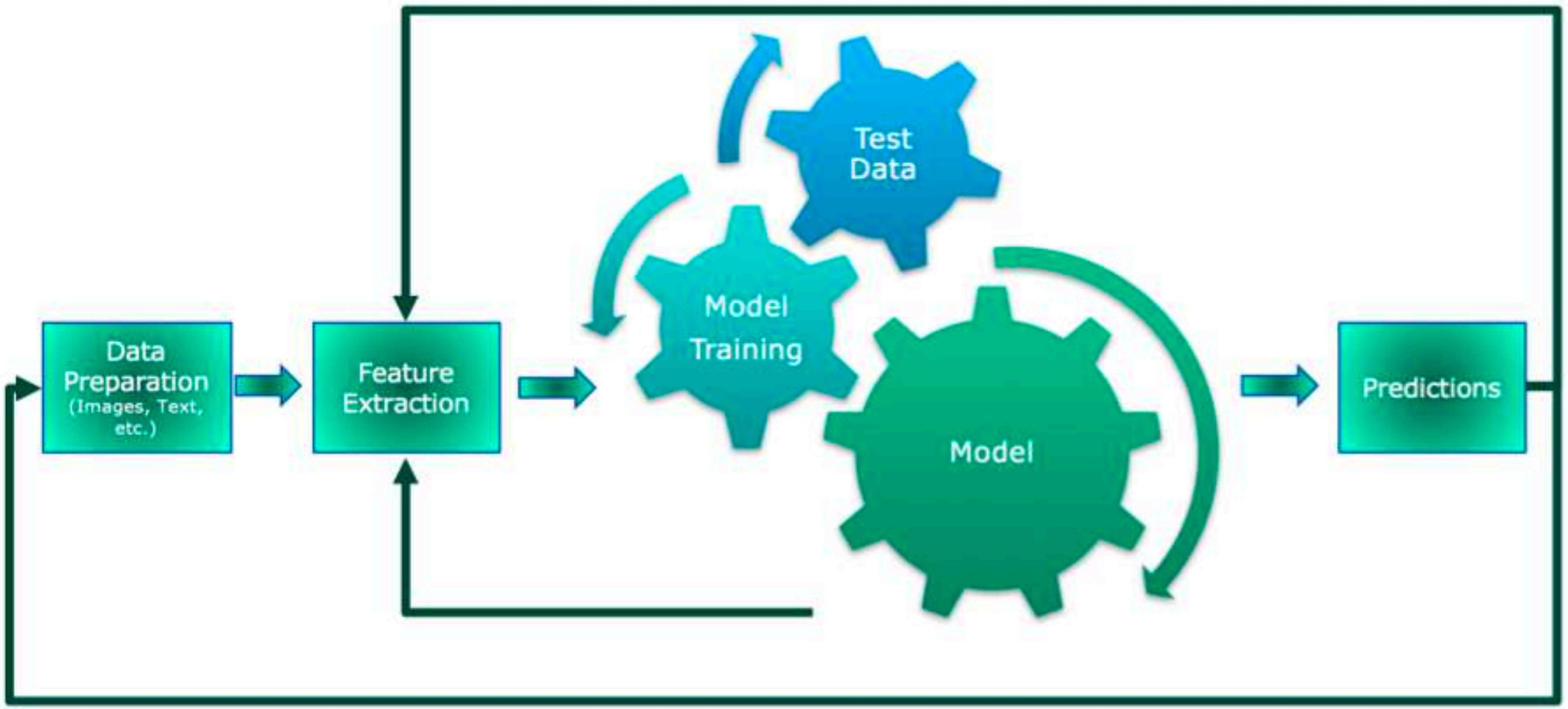
Ref: devopedia.org

NATURAL LANGUAGE PROCESSING (NLP)

- Statistical techniques to model text from a computational perspective
- Text mining: Objective (e.g. news events, financial events, terminology extraction) and subjective information extraction (e.g. sentiment analysis, emotion detection, personality, profiling) from text
- Lots of applications!



A Standard Machine Learning Pipeline



SUPERVISED MACHINE LEARNING

- AI field that studies computer algorithms for automatically learning complex properties from training data and make predictions for new data
- Classifier: supervised machine learning technique that performs classification:
  - Training data: each item is labeled with the correct class label
  - Test data: class label? Predicted based on the model learned on the basis of the training data

CLASSIFIER: TRAINING

Data	Class label	Features
	fish	Animal, jaws, black, orange, white
	panther	Animal, paws, black, black, black
	bird	Animal, feathers, black, orange, white

GHENT UNIVERSITY

CLASSIFIER: TEST

Data

Features

Animal, feathers, black,
orange, black

Class label

bird

FEATURES

- Relevant information to solve the task
- features for NLP:
  - Lexical features (e.g. words in a sentence, dictionary lookup)
  - Semantic features
  - Grammatical/syntactic features
  - …

GHENT UNIVERSITY

CORPORA

- A corpus = a collection of computer-readable texts, E.g. the Brown corpus (Kučera and Francis, 1967) with ± 1M English words from different text genres, the Google N-gram corpus (Lin et al., 2012) with 1 trillion tokens of historical and (some) specialized texts

- Corpora are the basis of any NLP task, allowing to extract information, find and learn patterns, calculate n-gram frequencies and co-occurrences (see later), etc.

TEXT PROCESSING

- Text processing = converting text to a standardized form to use it for computational analysis
- Rich lexical variety in natural language > meaningful representation
- Linguistic preprocessing:
    - Sentence splitting
    - Tokenization
    - Word normalization (lemmatization, stemming, …)
    - PoS-tagging
    - (dependency) parsing
    - Named entity recognition

[LETS PREPROCESS: WWW.LT3.UGENT.BE/LETS-DEMO/](http://www.lt3.ugent.be/lets-demo/)

Input:

Coronavirus: environ 1.000 personnes en quarantaine dans une ville allemande

La mesure s’applique au district de Heinsberg, à la frontière néerlandaise et à quelques kilomètres de la frontière belge

GHENT UNIVERSITY

FEATURES: N-GRAMS

- N-grams are sequences of *N* units/tokens.
- These units/tokens can be **characters, syllables, words** (including abbreviations, numbers, punctuation marks, emoji,...), **phonemes**, etc. depending on the application.
- *N* refers to the **size of the sequence**, e.g, 1-gram/unigram, 2-gram/bigram, 3-gram/trigram, 4-gram/tetragram,...

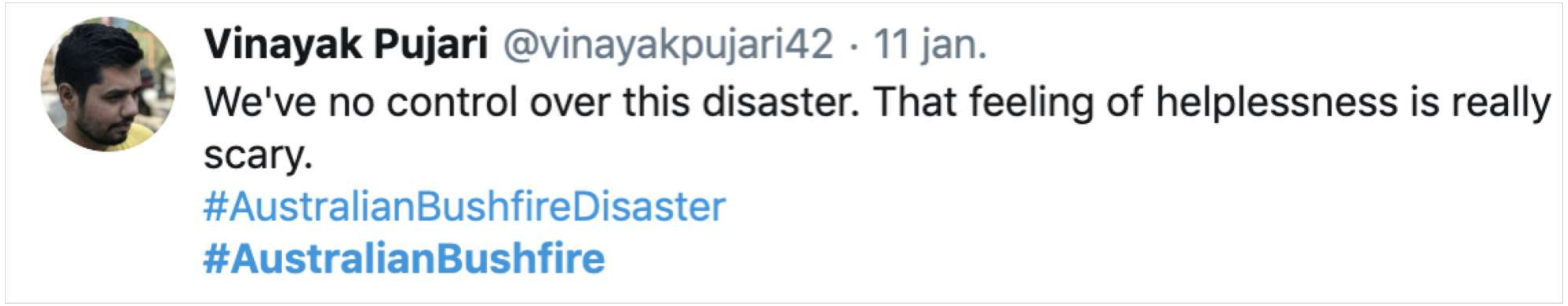
I love NLP.

→ Unigrams: “I”, “love”, “NLP”, “.”
→ Bigrams: “I love”, “love NLP”, “NLP .”

GHENT UNIVERSITY

N-GRAMS

- Basic units for many **NLP tasks**: syntactic parsing, PoS-tagging, classification, language modeling, machine translation, readability prediction, etc.
- Example: how *n*-grams can be useful for tasks like **sentiment analysis**:



“no control”
“disaster”
“helplessness”
“scary”
…

“WINNER”
“outstanding”
“awards”
…

GHENT UNIVERSITY

VECTOR SEMANTICS

VECTOR SEMANTICS

Distributional hypothesis*:

"words that occur in similar contexts tend to have similar meanings"

↓

The amount of meaning difference between two words corresponds roughly to the amount of difference in their environments (co-occurring words)

*Harris (1954), Firth (1957)

VECTOR SEMANTICS

What the "Green Deal" doesn't mention is that, as helpful as AI can, indeed, be in dealing with the dawning climate catastrophe, the AI industry itself has a rapidly growing ***carbon footprint***. Some researchers estimate that by 2040, the whole tech industry could contribute as much as 14 percent to the world's entire *carbon footprint*. And the AI industry, whose energy consumption has doubled during the last four years, plays a significant role in that: A powerful AI system that processes natural language, for example, omits 300,000 kilograms of *carbon* dioxide emissions while being trained, according to a study from earlier this year. That's about as much as 125 round-trip flights from New York City to Beijing.

Proposed tools such as a *carbon* border tax — EU tariffs on imported goods based on their ***CO2 footprint*** — could be seen as a protectionist measure and a violation of World Trade Organization rules, for example.

Von der Leyen has insisted measures to make the bloc climate neutral are "a long-term economic imperative."

Besides the *carbon* border tax, ongoing efforts to boost the role of the euro in global transactions — including in energy payments — are also meant to help the bloc become the world's green growth champion and force others, especially economic competitors, to follow suit.

Cities consume more than two-thirds of the world 's energy, and account for more than 70 per cent of global *carbon* dioxide emissions. The choices that will be made on urban infrastructure in the coming decades on urban planning, energy efficiency, power generation and transport will have decisive influence on the emissions curve. Indeed, cities are where the climate battle will largely be won or lost.

But in addition to their enormous ***climate footprint***, cities generate more than 80 per cent of global gross domestic product and, as centers of education and entrepreneurship, they are hubs of innovation and creativity, with young people often taking the lead.

GHENT UNIVERSITY

VECTOR SEMANTICS

What the "Green Deal" doesn't mention is that, as helpful as AI can, indeed, be in dealing with the dawning climate catastrophe, the AI industry itself has a rapidly growing ***carbon footprint***. Some researchers estimate that by 2040, the whole tech industry could contribute as much as 14 percent to the world's entire *carbon footprint*. And the AI industry, whose energy consumption has doubled during the last four years, plays a significant role in that: A powerful AI system that processes natural language, for example, omits 300,000 kilograms of *carbon* dioxide emissions while being trained, according to a study from earlier this year. That's about as much as 125 round-trip flights from New York City to Beijing.

Proposed tools such as a *carbon* border tax — EU tariffs on imported goods based on their ***CO2 footprint*** — could be seen as a protectionist measure and a violation of World Trade Organization rules, for example.

Von der Leyen has insisted measures to make the bloc climate neutral are "a long-term economic imperative."

Besides the *carbon* border tax, ongoing efforts to boost the role of the euro in global transactions — including in energy payments — are also meant to help the bloc become the world's green growth champion and force others, especially economic competitors, to follow suit.

Cities consume more than two-thirds of the world 's energy, and account for more than 70 per cent of global *carbon* dioxide emissions. The choices that will be made on urban infrastructure in the coming decades on urban planning, energy efficiency, power generation and transport will have decisive influence on the emissions curve. Indeed, cities are where the climate battle will largely be won or lost.

But in addition to their enormous ***climate footprint***, cities generate more than 80 per cent of global gross domestic product and, as centers of education and entrepreneurship, they are hubs of innovation and creativity, with young people often taking the lead.

GHENT UNIVERSITY

VECTOR SEMANTICS

Vector semantics = learning representations of the meaning of words directly from their distributions in texts:

→ a word's distribution is the set of concepts in which it occurs, the neighboring words or grammatical environments

→ two words that occur in very similar distributions are likely to have the same meaning

GHENT UNIVERSITY

What does “bamkimuk” mean?

- He handed her a glass of red bamkimuk.
- Beef dishes are made to compliment this bamkimuk.
- He was feeling dizzy, because he drank too much bamkimuk.
- She drank some chilled white Californian bamkimuk with her bread and cheese.

bamkimuk = ??

What does “bamkimuk” mean?

- He handed her a glass of red bamkimuk.
- Beef dishes are made to compliment this bamkimuk.
- He was feeling dizzy, because he drank too much bamkimuk.
- She drank some chilled white Californian bamkimuk with her bread and cheese.

bamkimuk = wine

GHENT UNIVERSITY

Multilingual contextual approach

Words occurring in similar contexts tend to be semantically similar

⇩

If the source and target terms have similar contexts → translations

?

?

BREW

?

?

GHENT UNIVERSITY

Context of word = most frequent collocates in corpus

Orang Utan Brewery and Pub - Asia's first to	brew	its own ale - has opened, following problems
duties and other regulations. It is able to	brew	10 different types of beer, including exotically
become so popular that Guinness decided to	brew	in the country. In 1965, the first bottle
local brewers Multi Bintang Indonesia - who	brew	FES - as the starting block to success.
manager. and are pictured happily trying the	brew	with a not so happy who was unfortunately
transplanted into the wrong soil. While she went to	brew	coffee Fletcher introduced Patrick and
a surprise,' I smiled. 'A cheeky little	brew	but you'll be amused by its pretension.
the table and sipped at the dark, bitter	brew	self-consciously, aware of his eyes, cool
ozone, iodine and women. What a tempting	brew	. I've not enjoyed a dip in the briny since
gin-aholic young females to come and sample our	brew	. We started it yesterday, so it should
and dying animals make a strong political	brew	. America, which has not had a big oil spill
Tom, 'Going To Nepal', a heady guitar/pop	brew	, was inspired by a real life experience
this has yet to happen. Charlotte	Brew	was the first woman to ride in the National
But put together, they produced a potent	brew	which has caused an extraordinary about-face
and then to reggae, the resulting musical	brew	was explosive. 'There can be no liberation

GHENT UNIVERSITY

Look up contexts for terms

English terms

brew

beer ale bottle coffee bitter dark lager guinness ginger

Japanese terms

醸造

ビール ボトル 苦い ラガー 滴ります

ドリンク

水 ガラス ビール 煙 シャン

します

水 土壌 ミルク ビール 廃棄物

➔ How to compare contexts in different languages?

GHENT UNIVERSITY

Translation lexicon:translate context words

English terms

brew

beer ale bottle coffee
bitter dark lager
guinness ginger

Japanese terms

醸造

ビール ボトル
苦い ラガー
滴ります

ドリンク

水 ガラス
ビール 煙
シャン

します

水 土壌
ミルク ビール
廃棄物

Translation lexicon

water 水
glass ガラス
beer ビール

➔ Now we can compare the English and Japanese context words for all terms

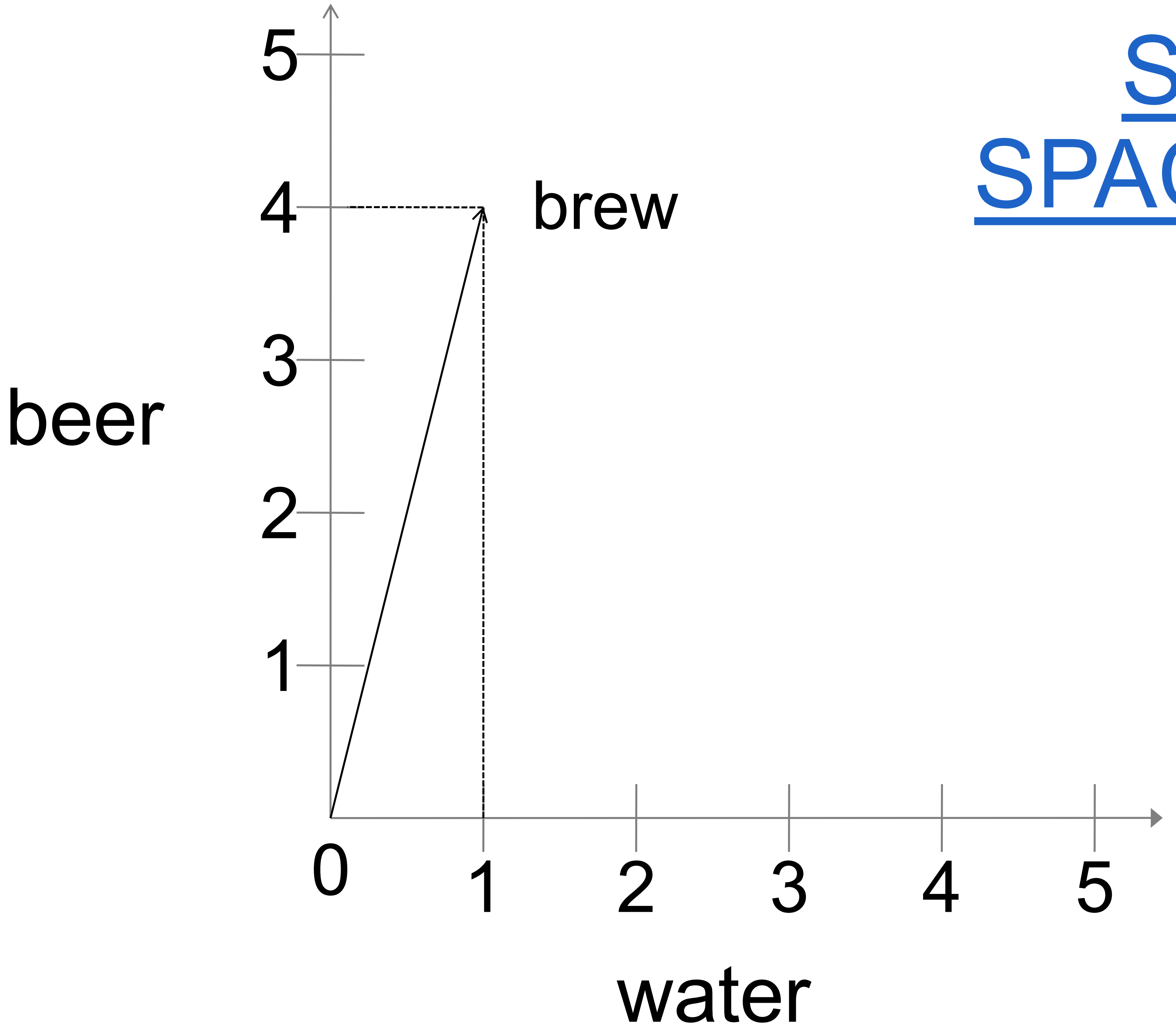
HOW TO COMPARE THESE CONTEXTS?

ENGLISH	beer	water
brew	4	1

JAPANESE	beer	water
醸造	5	1
ドリンク	2	4
します	1	4

GHENT UNIVERSITY

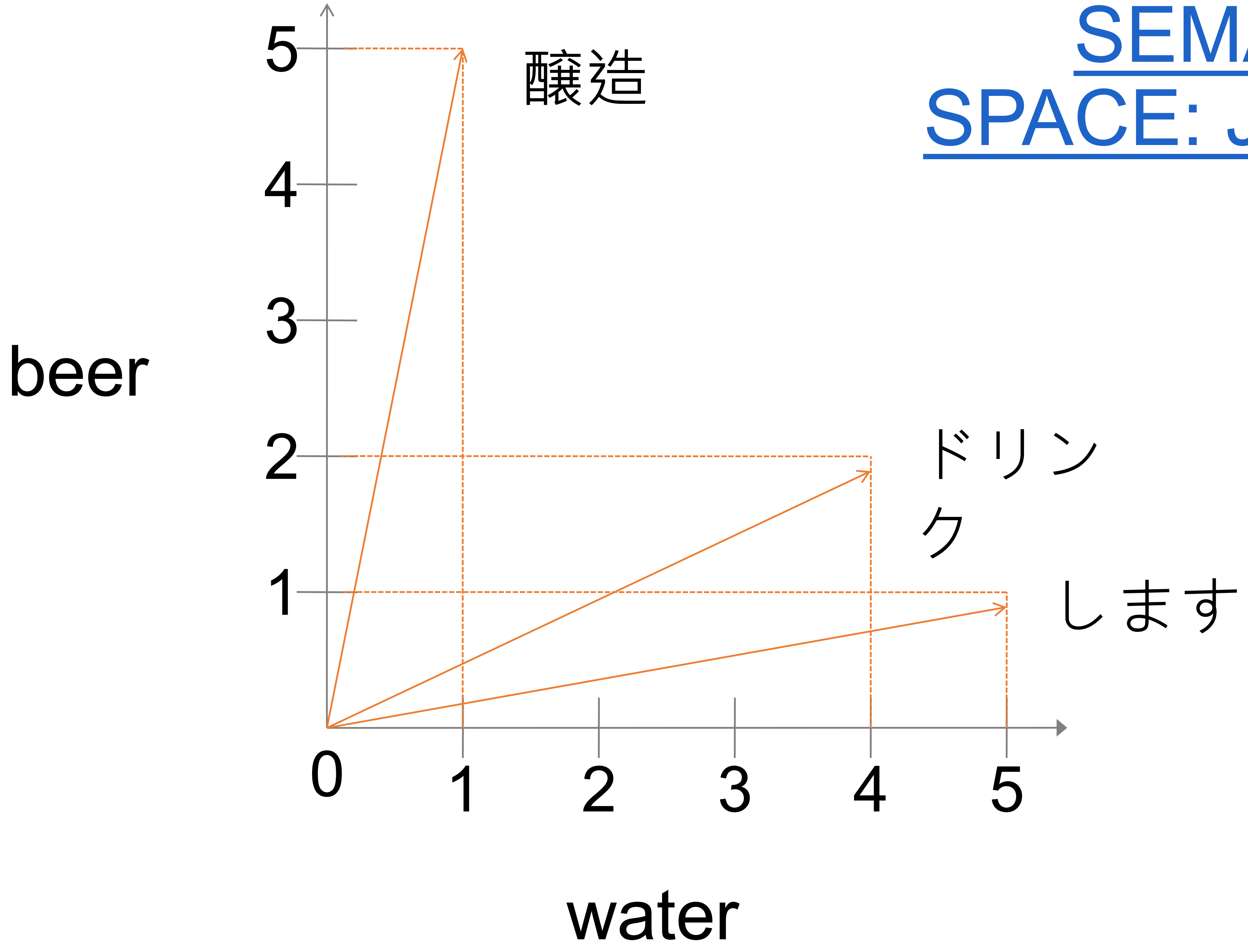
SEMANTIC SPACE: ENGLISH



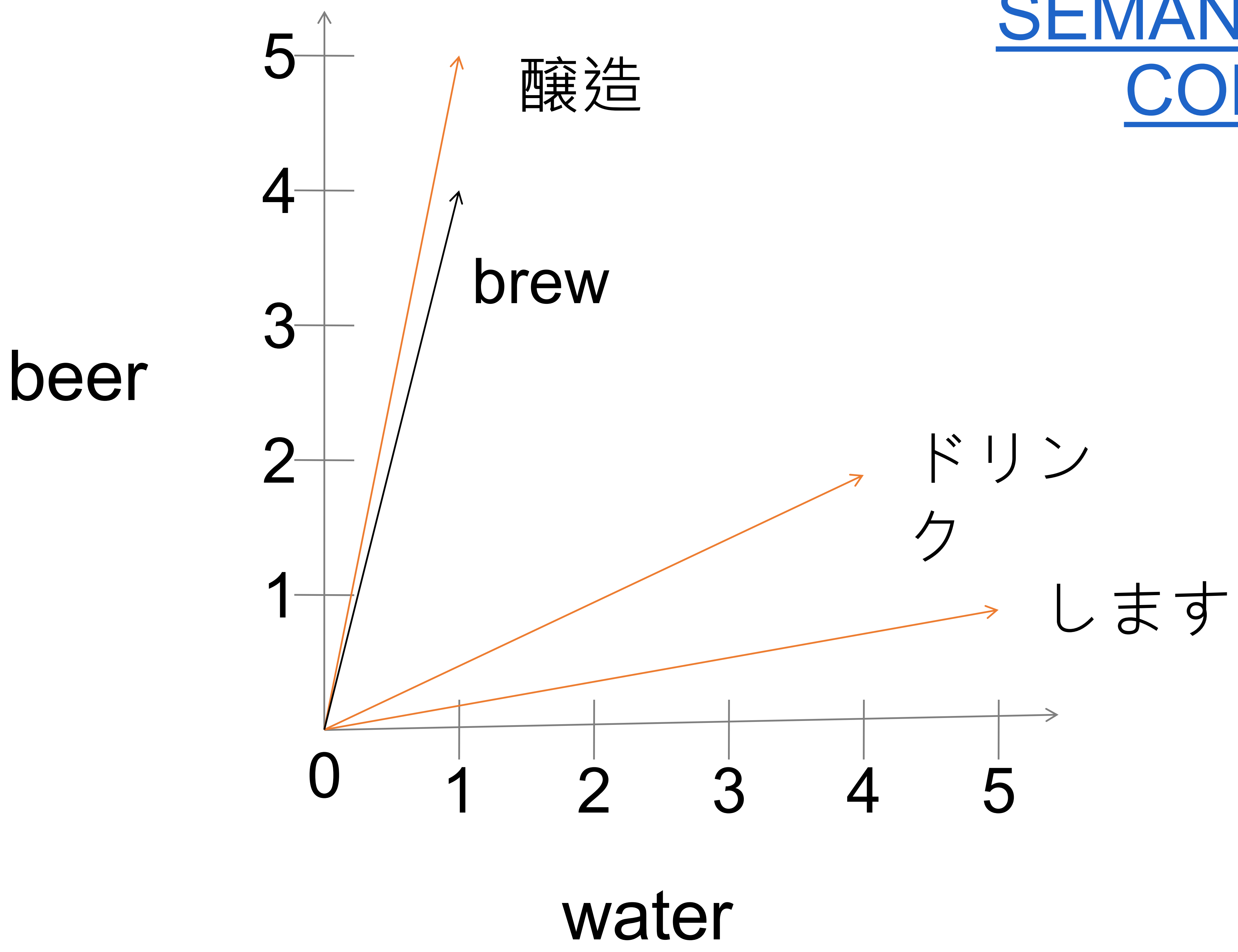
Ref: Turney & Pantel, 2010

GHENT UNIVERSITY

SEMANTIC SPACE: JAPANESE

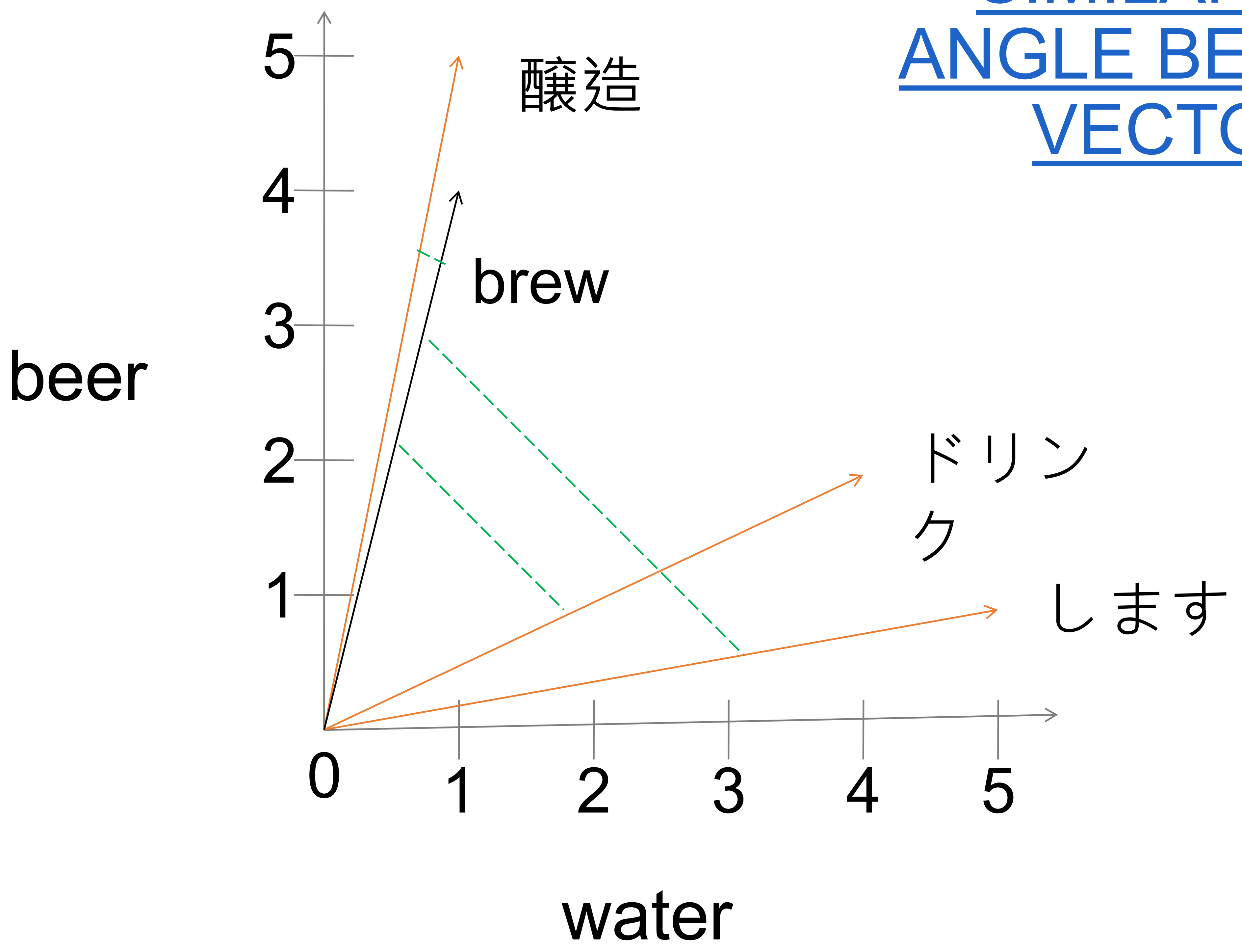


SEMANTIC SPACE: COMBINED



GHENT UNIVERSITY

SIMILARITY = ANGLE BETWEEN VECTORS



WORDS AND VECTORS

A term-term or term-context matrix represents words and their co-occurrence with other words (i.e. their context)

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

Figure 6.5 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

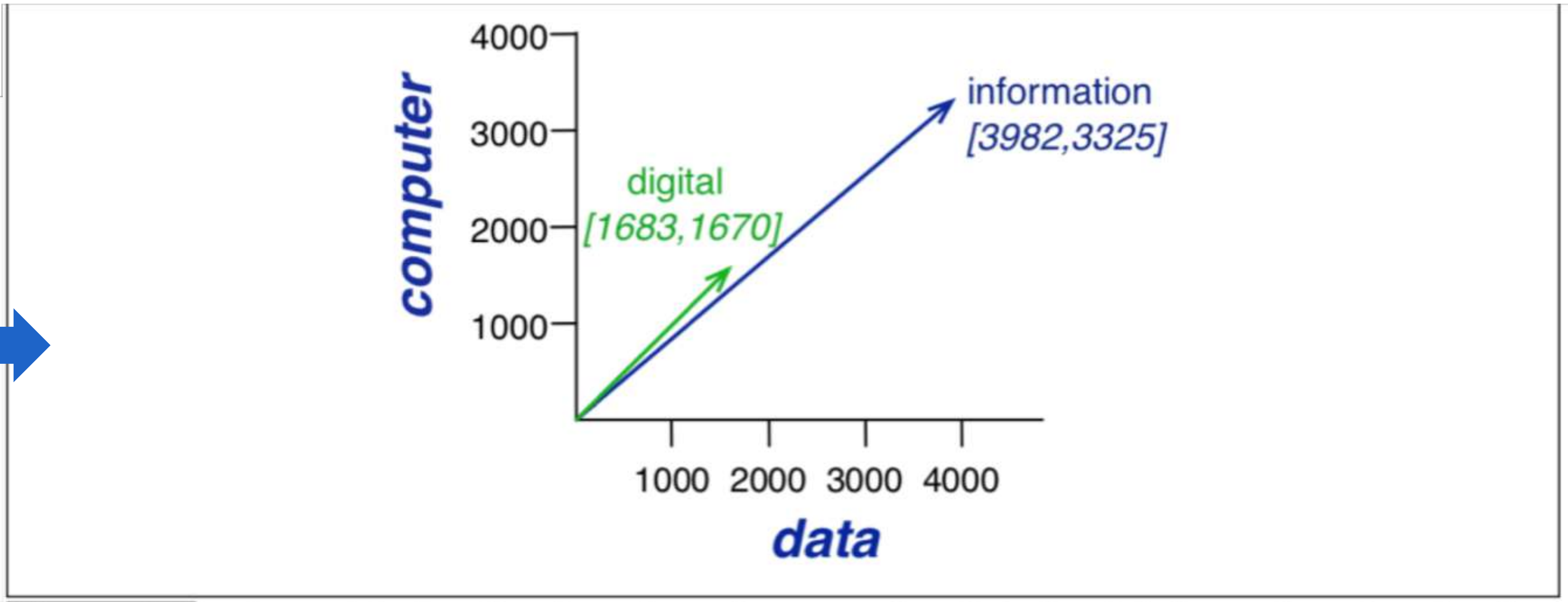


Figure 6.6 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *computer*.

→ the rows (*"digital"*, *"information"*) are **vectors**

→ the columns (*"computer"*, *"data"*) are **dimensions**

GHENT UNIVERSITY

WORD2VEC

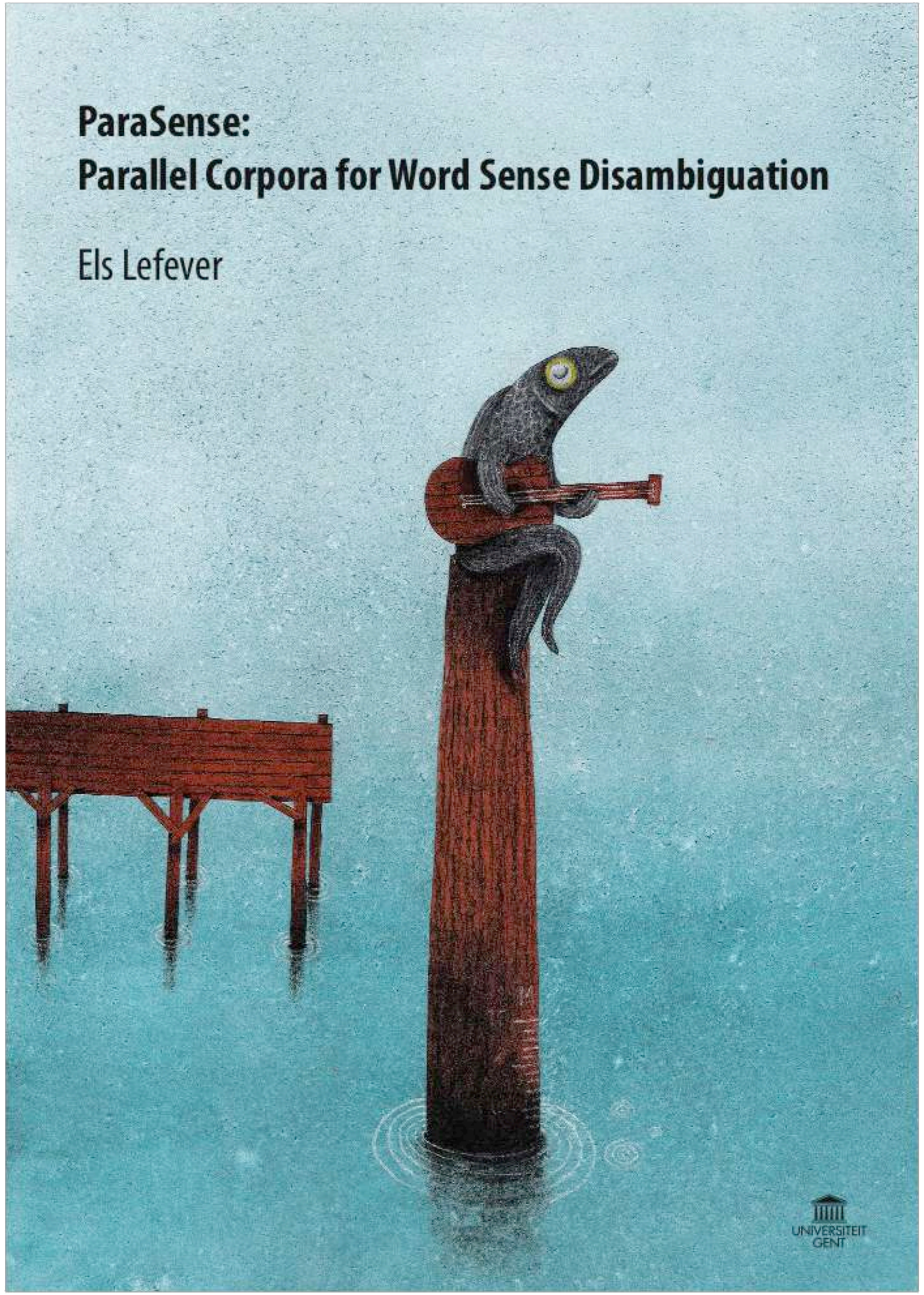
- Mikolov proposed to learn word vectors using a neural network with a single hidden layer (Mikolov et al 2013) => word2vec embeddings
- Intuition of the (skip-gram) algorithm: *“is word X likely to occur in the neighbourhood of word Y?”*
- Most important advantage of word2vec is that the algorithm learns the weights that make up the vectors based on raw input text
- Many neural architectures and models have been proposed for computing word vectors

  - GloVe (2014) - Global Vectors for Word Representation
  - FastText (2017) - Enriching Word Vectors with Subword Information
  - ELMo (2018) - Deep contextualized word representations
  - BERT (2019) - Bidirectional Encoder Representations from Transformers

USE CASES

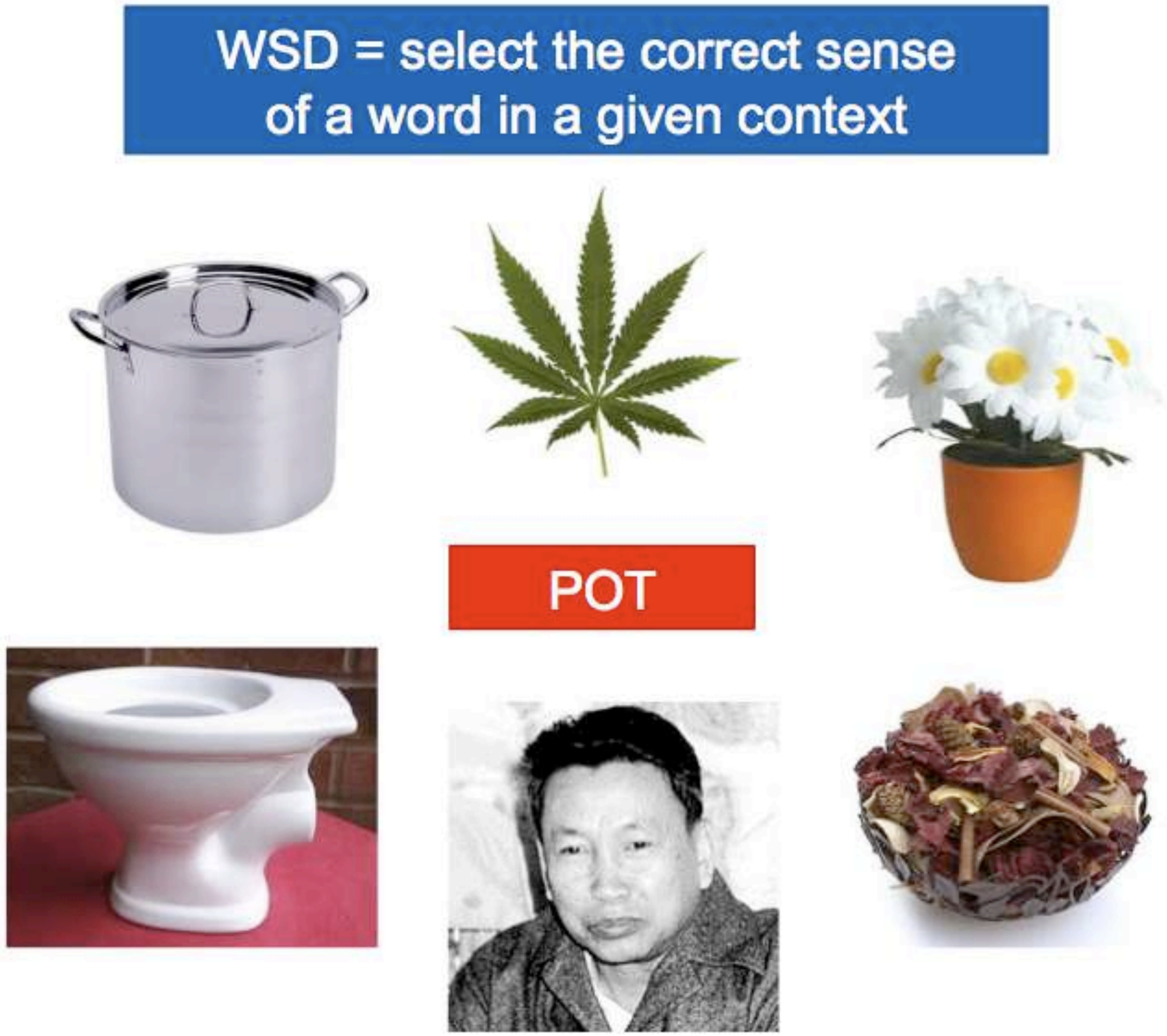
CROSS-LINGUAL WORD SENSE DISAMBIGUATION

GHENT UNIVERSITY



GHENT UNIVERSITY

WORD SENSE DISAMBIGUATION



e.g. WordNet labels:

http://wordnetweb.princeton.edu/perl/webwn

Noun

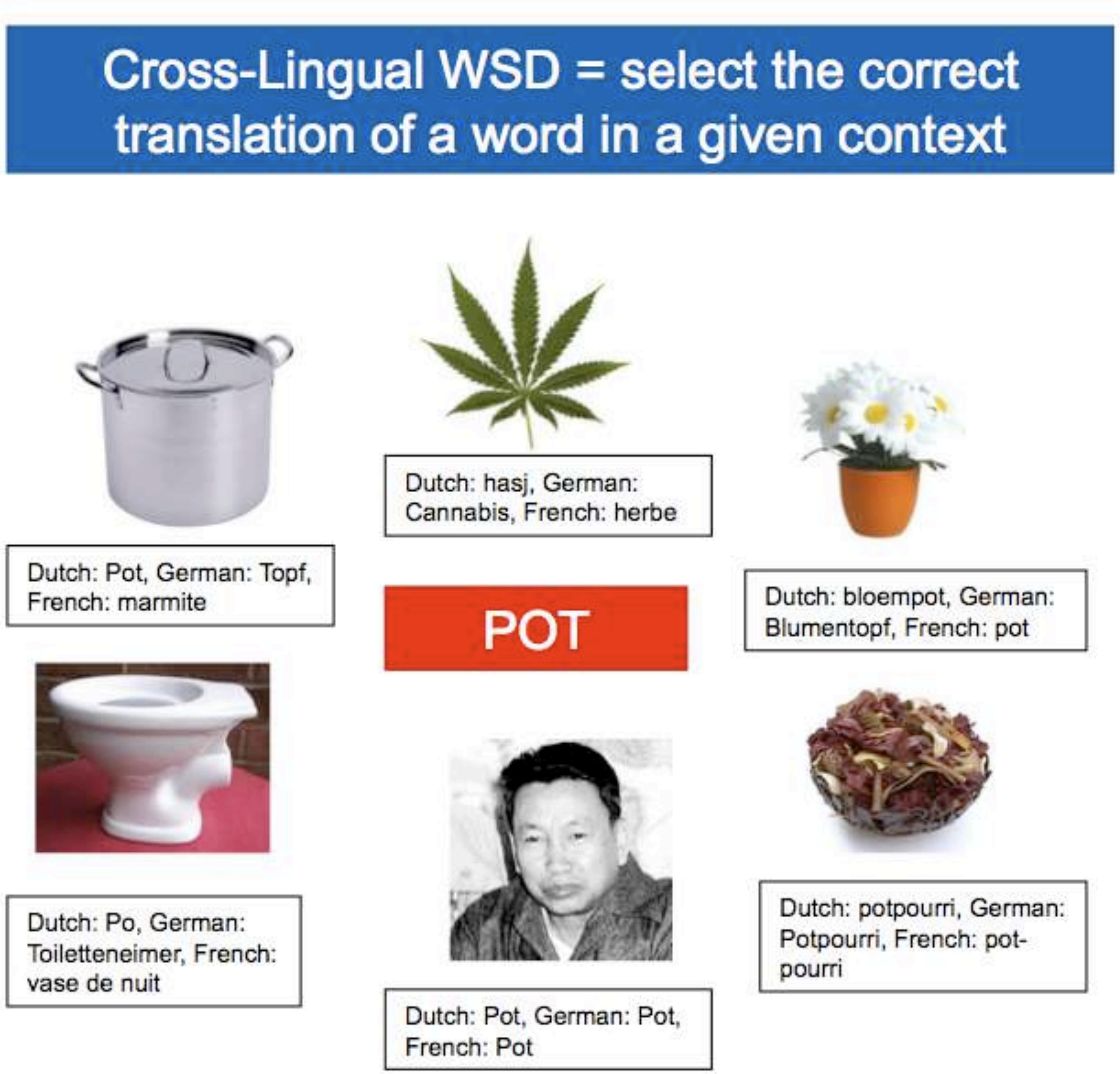
- S: (n) **pot** (metal or earthenware cooking vessel that is usually round and deep; often has a handle and lid)
- S: (n) toilet, can, commode, crapper, **pot**, potty, stool, throne (a plumbing fixture for defecation and urination)
- S: (n) **pot**, potful (the quantity contained in a pot)
- S: (n) **pot**, flowerpot (a container in which plants are cultivated)
- S: (n) batch, deal, flock, good deal, great deal, hatful, heap, lot, mass, mess, mickle, mint, mountain, muckle, passel, peck, pile, plenty, **pot**, quite a little, raft, sight, slew, spate, stack, tidy sum, wad ((often followed by `of') a large number or amount or extent) *"a batch of letters"; "a deal of trouble"; "a lot of money"; "he made a mint on the stock market"; "see the rest of the winners in our huge passel of photos"; "it must have cost plenty"; "a slew of journalists"; "a wad of money"*
- S: (n) **pot**, jackpot, kitty (the cumulative amount involved in a game (such as poker))
- S: (n) **pot**, potbelly, bay window, corporation, tummy (slang for a paunch)
- S: (n) potentiometer, **pot** (a resistor with three terminals, the third being an adjustable center terminal; used to adjust voltages in radios and TV sets)
- S: (n) **pot**, grass, green goddess, dope, weed, gage, sess, sens, smoke, skunk, locoweed, Mary Jane (street names for marijuana)

Verb

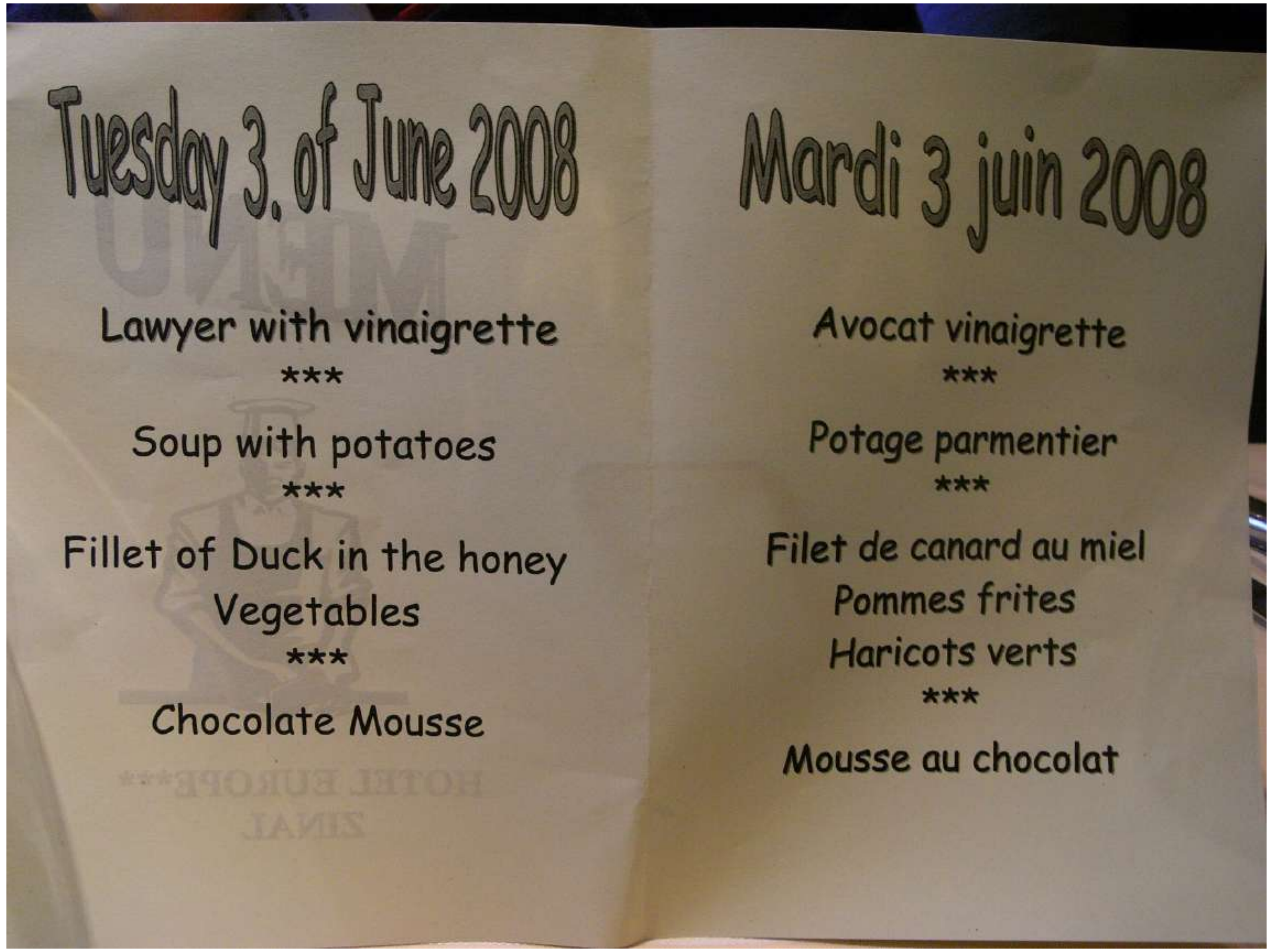
- S: (v) **pot** (plant in a pot) *"He potted the palm"*

GHENT
UNIVERSITY

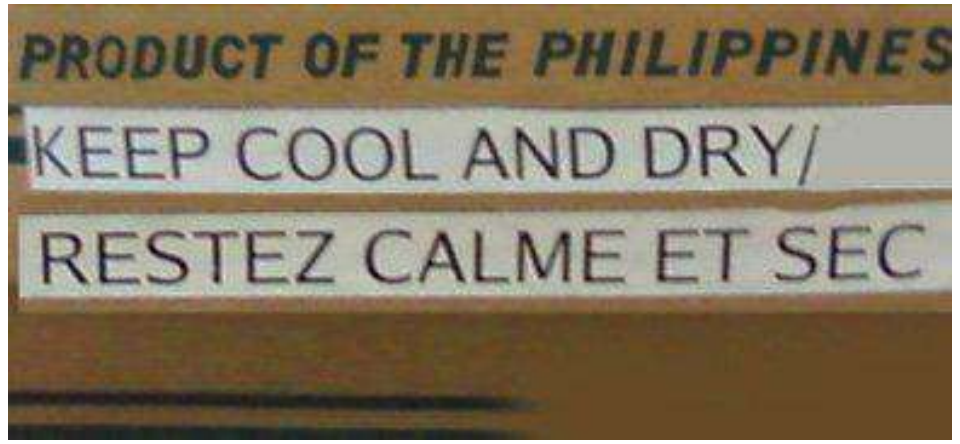
CROSS-LINGUAL WSD



USE?


Tuesday 3. of June 2008
Lawyer with vinaigrette
***
Soup with potatoes
***
Fillet of Duck in the honey
Vegetables
***
Chocolate Mousse
Mardi 3 juin 2008
Avocat vinaigrette
***
Potage parmentier
***
Filet de canard au miel
Pommes frites
Haricots verts
***
Mousse au chocolat


Windmill Sausage Co.Ltd.
"Where Sausage is our Speciality"
POLISH SAUSAGE
POLISSEZ LA SAUCISSE
KEEP FROZEN
GARDEZ CONGELÉ


PRODUCT OF THE PHILIPPINES
KEEP COOL AND DRY/
RESTEZ CALME ET SEC


Mixed Nuts
Écrous Mélangés
454 gr

CLASS LABEL

Example: *It is no longer the locomotive it once was, it is now the last coach in the train*

- Monolingual class label: coach%1:06:00
- Multilingual class label:
  wagon, Waggon, wagon, vagón, vagone

PARASENSE

- ParaSense = a truly multilingual classification-based machine learning approach to Word Sense Disambiguation.
- start from 2 basic assumptions:

1. possibility to use parallel corpora to extract translation labels and disambiguating information in an automated way

2. incorporating multilingual evidence will be more informative than monolingual or bilingual features

PREPROCESSING OF THE DATA

- Data: six-lingual sentence-aligned subcorpus of the Europarl parallel corpus containing one of the 20 ambiguous focus words (total: 35,686 sentences)

- Shallow Linguistic Analysis:
  - Tokenisation
  - Part-of-Speech tagging
  - Chunking
  - Lemmatisation

FEATURE VECTOR CONSTRUCTION

- combination of English local context features and a set of bag-of-words (ngram) translation features
- class labels: automatically generated word alignments for the ambiguous focus words

GHENT UNIVERSITY

LOCAL CONTEXT FEATURES

- features related to the focus word itself: word form, lemma, Part-of-Speech, chunk info
- local context features related to a 7-word window containing the ambiguous word
- Example: *It is no longer the locomotive it once was, it is now the last coach in the train*
  - features focus word: coach coach NN I-NP
  - features context word -3: now now RB I-ADVP
  - features context word -2: the the DT I-NP
  - features context word -1: last last JJ I-NP
  - features context word +1: in in IN I-PP
  - features context word +2: the the DT I-NP
  - features context word +3: train train NN I-NP

TRANSLATION FEATURES

a set of binary bag-of-words features from the aligned translations (four languages):

- **PoS-tagging** and **lemmatisation** on all aligned translations
- per ambiguous focus word, a **list of content words** (nouns, adjectives, verbs and adverbs) was extracted
- **one binary feature** per selected content word

TRANSLATION FEATURES

English

- *Sentence 1*: Our Europe, that melting **pot** of cultures, languages and people, is possible thanks to free movement and study programmes.
- *Sentence 2*: Macao, as has already been said, has always been a melting **pot** of cultures and of new meetings of cultures, of religions too, and has always been a territory where peace, tranquillity and coexistence between peoples of the most diverse ethnic backgrounds have reigned.

Italian

- *Sentence 1*: La nostra Europa, quel **crogiolo** di culture, lingue e persone, è possibile grazie alla libera circolazione e ai programmi di studio.
- *Sentence 2*: Macao, come è stato detto, è sempre stata un **crogiolo** di culture, civiltà e religioni, una regione in cui le etnie più diverse convivono in pace e serenità.

TRANSLATION FEATURES

Italian

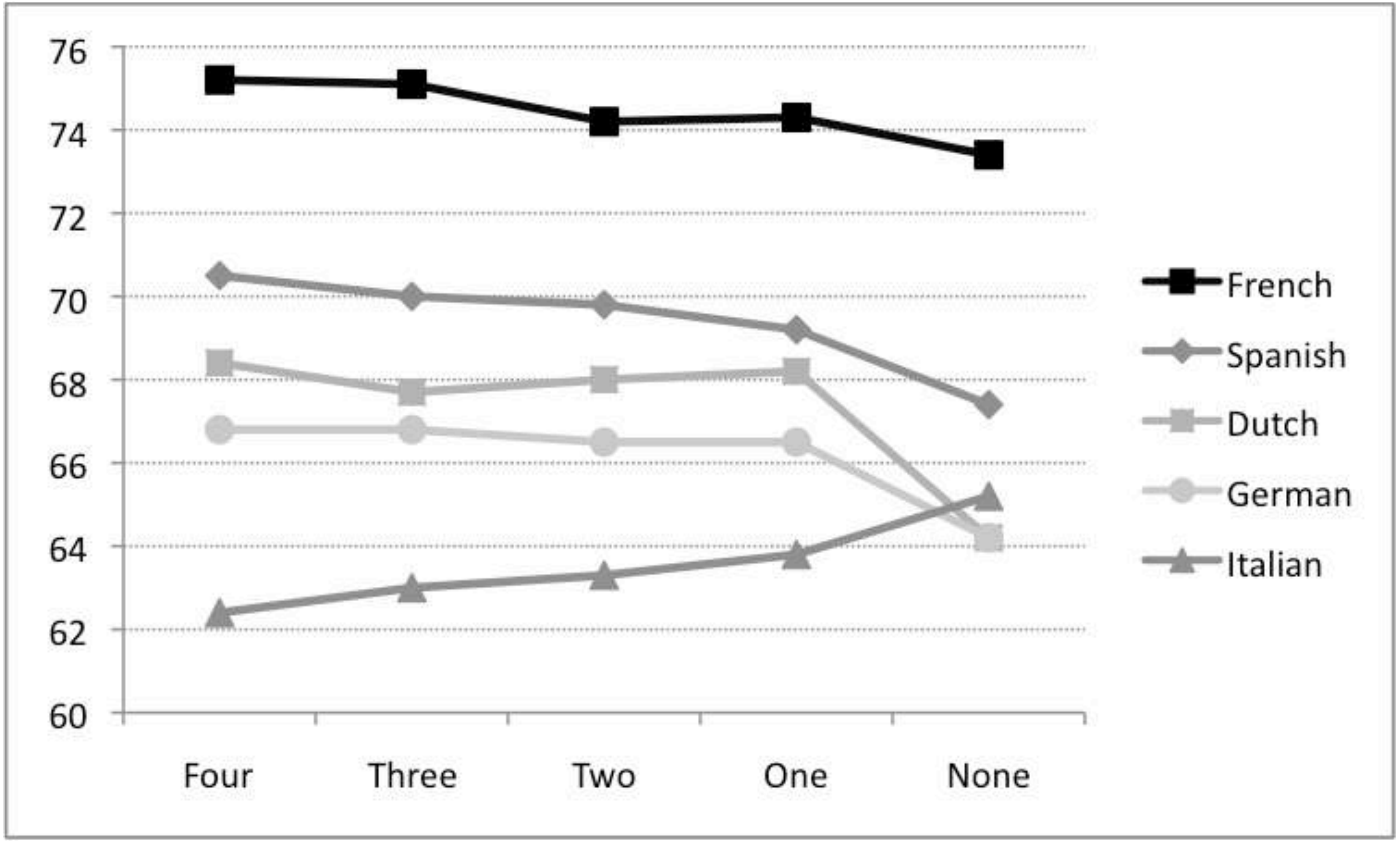
• *Sentence 1*: La nostra Europa, quel **crogiolo** di culture, lingue e persone, è possibile grazie alla libera circolazione e ai programmi di studio.

• *Sentence 2*: Macao, come è stato detto, è sempre stata un **crogiolo** di culture, civiltà e religioni, una regione in cui le etnie più diverse convivono in pace e serenità.

	Europa crogiolo cultura lingua persona essere possibile grazie libero circolazione programma studio Macao dire sempre civiltà religione regione etnia più diverso convivere pace serenità
Sentence 1	1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
Sentence 2	0 1 1 0 0 1 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1

GHENT UNIVERSITY

RESULTS



GHENT UNIVERSITY

CYBERBULLYING DETECTION

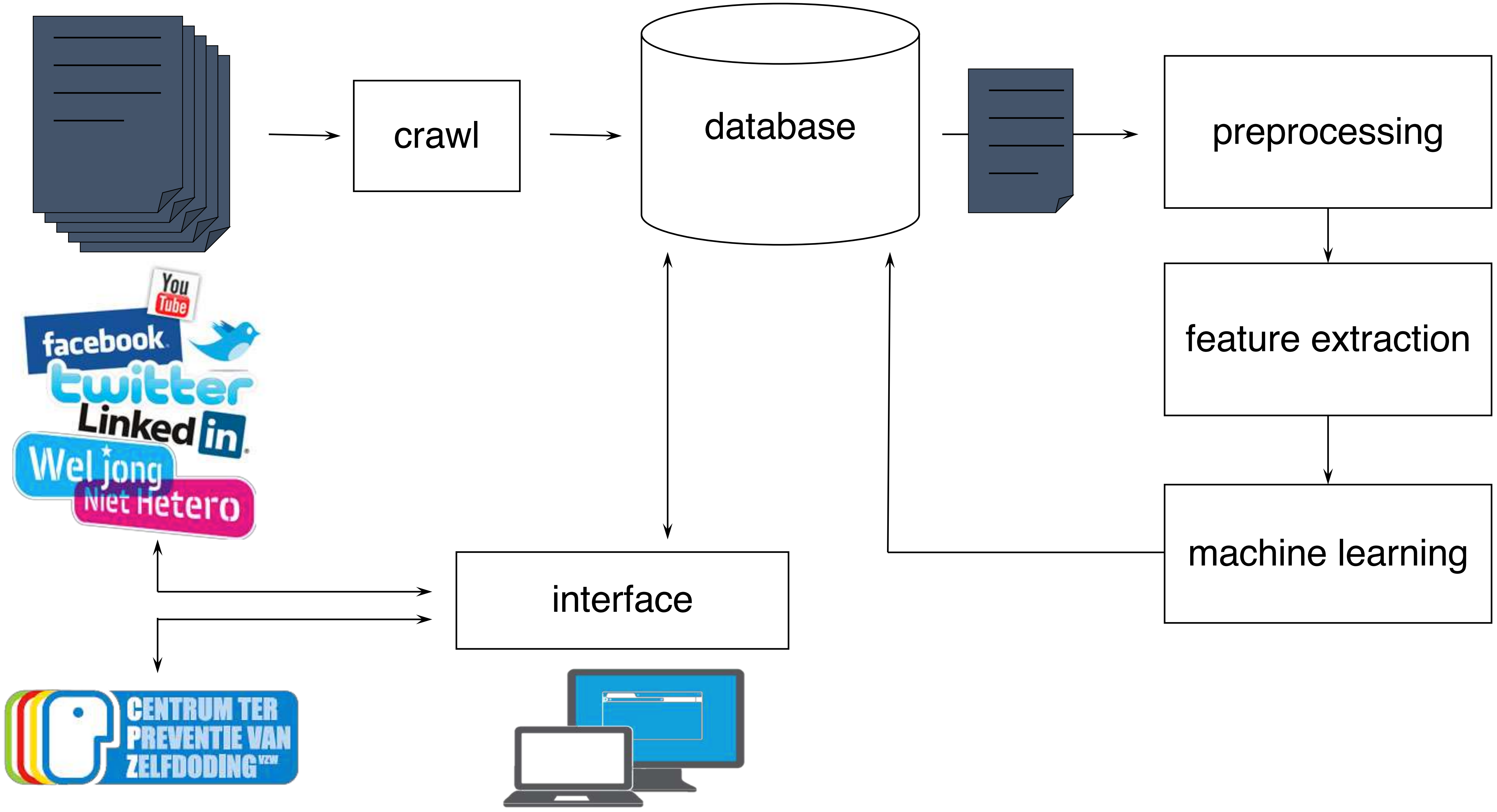
AMICA

- Detect situations that are harmful or threatening to young people in social networks
  - Cyberbullying
  - Sexually transgressive behaviour (for example grooming by paedophiles)
  - Depression and suicide announcement

=> Facilitate efficient action by moderators, police, parents, peer group, social services

GHENT UNIVERSITY

WORKFLOW


crawl
database
preprocessing
feature extraction
machine learning
interface
You Tube
facebook
twitter
Linked in
Wel jong
Niet Hetero
CENTRUM TER
PREVENTIE VAN
ZELFDODING VZW

PREPROCESSING / NORMALISATION OF USER-GENERATED TEXT

GHENT
UNIVERSITY

USER GENERATED CONTENT

Social media: blogs and microblogs (Twitter: 190 million tweets/day), wikis, podcasts, social networks (Facebook: 70 billion shares/month)

⇒Enormous amount of UGC

GHENT UNIVERSITY

UGC NORMALIZATION

Maxims of chat language:

- Write as fast as you can (fluent interaction)
  - Abbreviations, letter omission, acronyms, flooding, concatenation, capitalization, punctuation, spelling and grammar errors, …
- Write as you speak (informal character of the conversation)
  - Dialectical, phonetic, emoticons, …

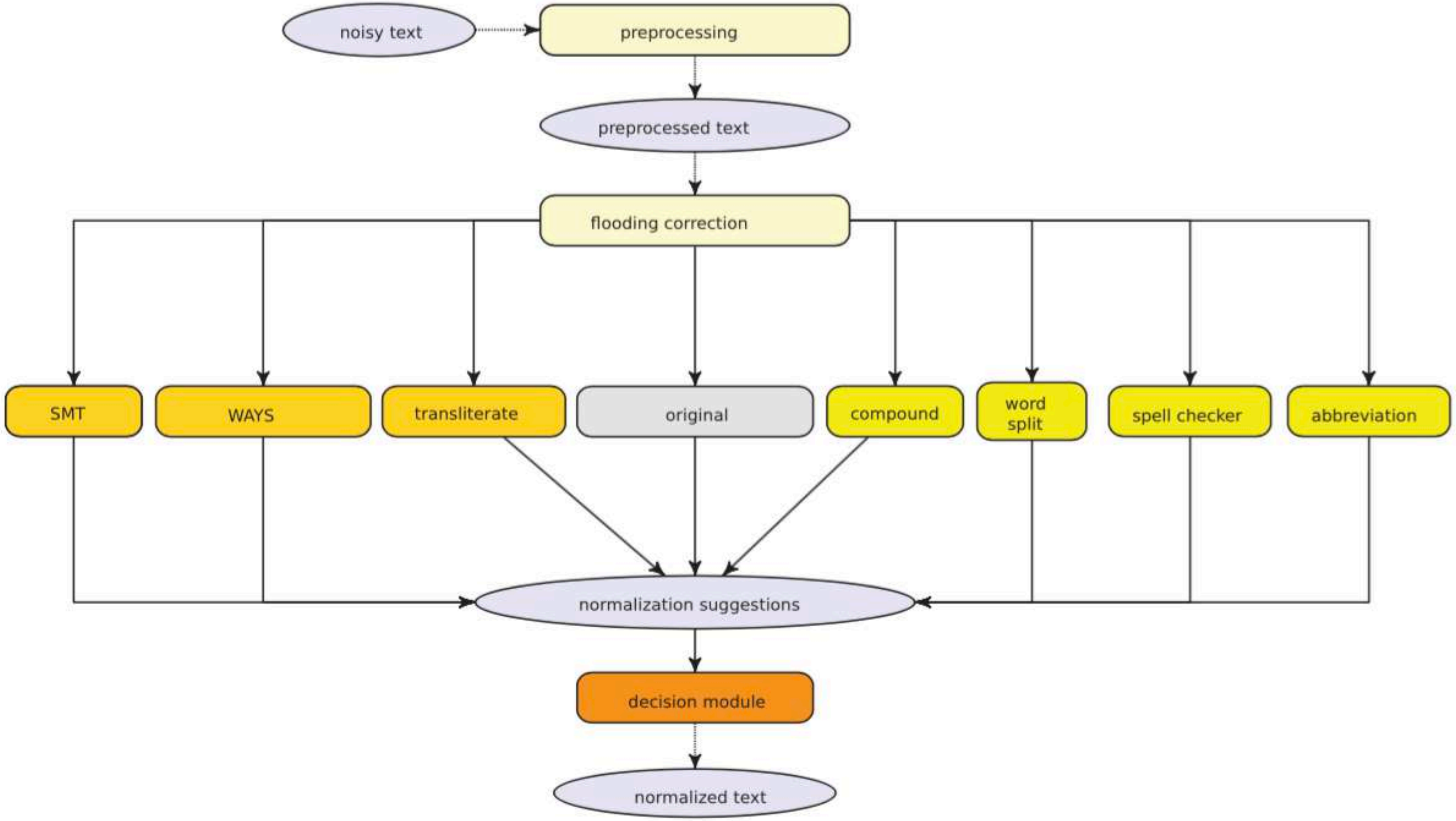
PROPERTIES OF CHAT LANGUAGE

- Omission of words / characters (spoke – spoken)
- Abbreviations, acronyms (LOL – laughing out loud)
- Deviations from standard spelling (luv – love, you iz – you are)
- Expression of emotions:
  - Flooding (looooooooove)
  - Emoticons (:p)
  - Capitalized letters (STUPID)
- Dutch-specific:
  - Concatenation of tokens (khou – ik hou)
  - Elimination of clitics and pronouns (edde – heb je)
  - Lot of dialects!

PROBLEM FOR TEXT ANALYSIS TOOLS

- Most NLP tools are developed for or trained on standard language
- They fail miserably on UGC
- Solutions
  - Develop new tools
    - E.g. Tweet NLP (CMU): http://www.cs.cmu.edu/~ark/TweetNLP/
  - Normalize the ‘non-standard’ language

ENSEMBLE APPROACH



GHENT UNIVERSITY

Reference: Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, and Lieve Macken. 2016. Multimodular text normalization of Dutch user-generated content. *ACM Trans. Intell. Syst. Technol.* 7, 4, (July 2016)

MODULES

- Preprocessing
  - Tokenization and sentence splitting
    - includes emoticons, emojis etc.
  - Character floooooooding

- Token-based modules
  - Abbreviations
    - Expansion dictionary (~ 350 abbrevs)
  - Spell checker
    - Levenshtein on dictionary (~ 2.3 million words)
  - Compound Module
    - Checks if a pair of words is actually one word
  - Word Splitter
    - ‘misje’ = ‘mis je’ (miss you)

GHENT UNIVERSITY

MODULES

- Context-based modules
  - Statistical Machine Translation
    - Token-unigram, character unigram, character-bigram and combinations
  - Transliteration (supervised ML)
    - supervised ML, memory-based learning style
      - +da+_n **i** ++_ged -> **iet**
  - WAYS (Write As You Speak): G2P + P2G (memory-based learning)
    - ni (niet, *not*)
    - kem (ik heb, *I have*)

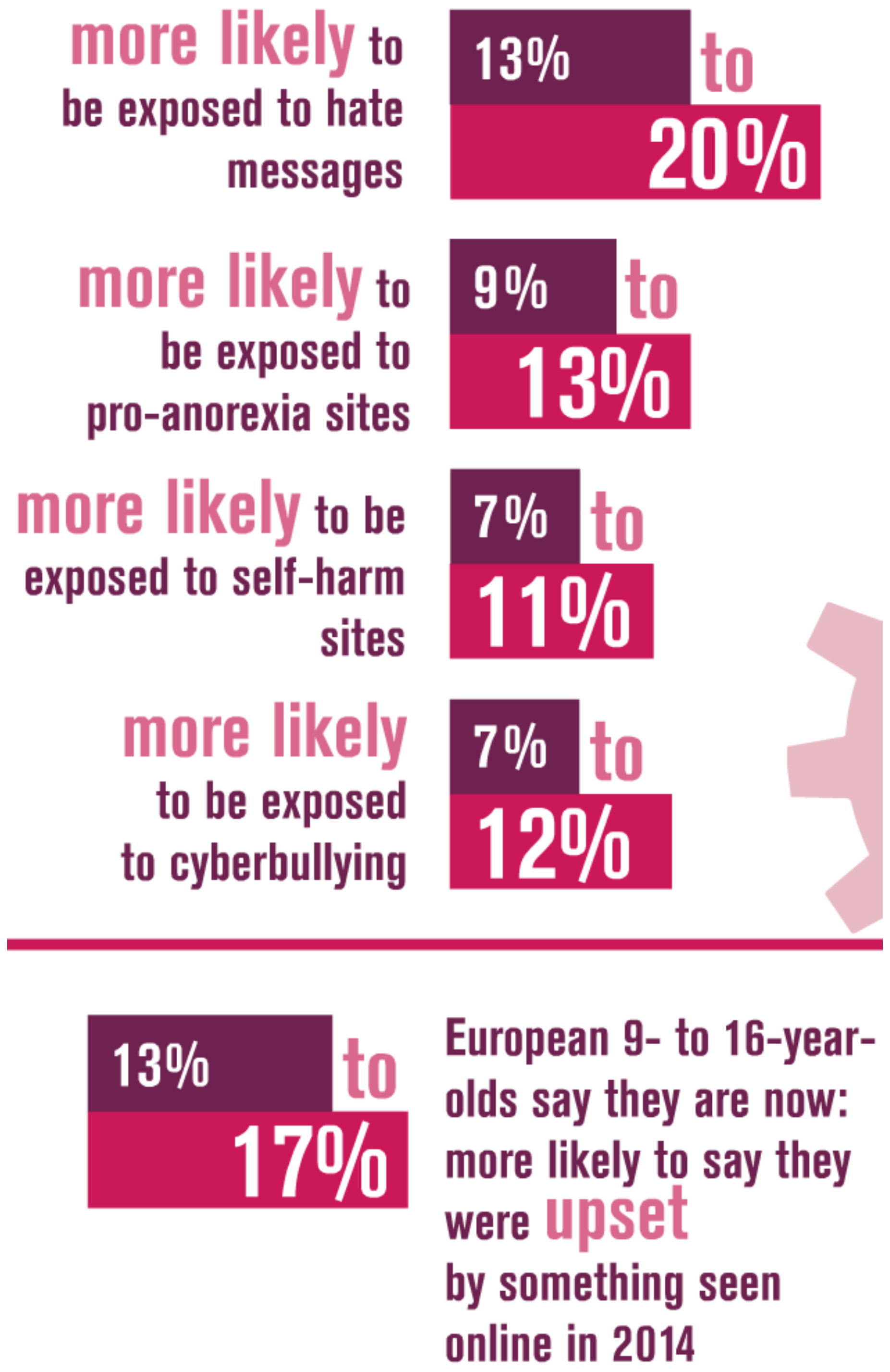
- "Original" Module
  - Many words are correct

GHENT UNIVERSITY

USE CASE: CYBERBULLYING DETECTION

RESEARCH MOTIVATION

- ± 20-40% of all youth have been victimized online (Tokunaga, 2010)
- Anonymity, lack of supervision and impact make social media a convenient way for cyberbullies to target their victim (Hinduja & Patchin, 2006)
- Information overload on the Web has made manual monitoring unfeasible



Source: the EU Kids Online report (2015)
http://www.lse.ac.uk/media@lse/research/EUKidsOnline

GHENT UNIVERSITY

DATA SET CONSTRUCTION

- We need large data sets to train machine learning systems
- Data collection for Dutch and English
  - Data from relevant social media
  - BUT: few / private data
  - Media campaign for donating examples of cyberbullying messages
  - BUT: sensitive data!
  - Cyberbullying simulations

GHENT UNIVERSITY

DATASET CONSTRUCTION: SIMULATION EXPERIMENTS

- Role playing in secondary schools on social media platform: FB-like social network, scenarios, profile cards (roles), debriefing
- Additional goal: education (prevention)

GHENT UNIVERSITY

DATA ANNOTATION

- Brat rapid annotation tool (Stenetorp et al., 2012)
- Two annotation levels (Van Hee et al., 2015)
  - Post level
    - Cyberbullying -vs- non-cyberbullying

      textual content that is published online by an individual and that is aggressive or hurtful against a victim.
    - Harmfulness score
      - 0 → the post does <u>not</u> contain indications of cyberbullying
      - 1 → the post contains <u>indications</u> of cyberbullying, although they are <u>not severe</u>
      - 2 → the post contains <u>serious indications</u> of cyberbullying
    - Author's role
      - Harasser
      - Victim
      - Bystander-defender
      - Bystander-assistant

GHENT UNIVERSITY

DATA ANNOTATION

- (Sub)sentence level: identification of fine-grained text categories related to cyberbullying
  - Threat/blackmail
  - Insult
  - Curse/exclusion
  - Defamation
  - Sexual talk
  - Defense
  - Encouragements (to the harasser)

Reference: Guidelines for the fine-grained analysis of cyberbullying, version 1.0 (2015)
Van Hee, C., Verhoeven, B., Lefever, E., De Pauw, G., Daelemans, W., & Hoste, V.

DATA ANNOTATION

Category	Brat annotation example	Translation
Threat/blackmail Expressions containing physical or psychological threats, or indications of blackmail.	1_Har · Threat or Blackmail · 2_Har s ik u tegen kom zieke rak op u gezicht x ¶	*I'll smash you in the face when I see you x*
Insult Expressions containing abusive, degrading or offensive language that are meant to insult the addressee.	1_Har · General insult · General insult · ¶ HAHAHAHA LOSER GIJ:(X AARDAPPELKOP	*HAHAHAHA YOU LOSER :(X POTATO HEAD*
Curse/exclusion Expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group.	2_Har · Curse or Exclusion · General insult · ¶ Pleeg zelfmoord niemand vindt u geestig ...	*Just commit suicide, nobody thinks you're funny...*
Defamation Expressions that reveal confident, embarrassing or defamatory information about the victim to a large public.	1_Har · Defamation · ¶ u mama versiert andere mannen hahahaha	*Your mom is flirting with other men hahahaha*
Sexual talk Expressions with a sexual meaning that are possibly harmful.	1_Har · Sexual harassment · ¶ Stuur my u naaktfoto, nu!!	*Send me a naked picture of yourself, now!!*
Defense Expressions in support of the victim, expressed by the victim himself or by a bystander.	1_Bystander_defender · General victim defense · General victim defense · ¶ Meid, koppie omhoog he! Laat je ni doen door die domme anoniempjes	*Cheer up girl, don't let those stupid anons make you feel bad*
Encouragements to the harasser Expressions in support of the harasser.	2_Bystander_assistant · General insult · Encouraging harasser · ¶ inderdaad ze is geen leven waard !!	*Indeed, she shouldn't be alive !!*

GHENT UNIVERSI

CYBERBULLYING EXPERIMENTS

- Class
  - Binary (bullying or non-bullying)
  - Binary (for each fine-grained class)
- Features
  - Word unigrams and bigrams
  - Character trigrams
  - Subjectivity lexicon features
  - Lexicon features (diminishers, intensifiers, proper names, negation words)
  - topic model features
- Classifier: SVM (Pattern) with linear kernel
- Data: ~85,000 posts
- Annotation agreement (kappa) 60-65%
- Very skewed data, scarce positive data (~10%)

GHENT UNIVERSITY

Reference: Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., et al. (2018). Automatic detection of cyberbullying in social media text . (H. Suleman, Ed.)*PLOS ONE*, *13*(10).

RESULTS BULLYING /VS/ NON-BULLYING

	Precision	recall	F1-score
EN	73%	57%	64%
NL	71%	53%	61%

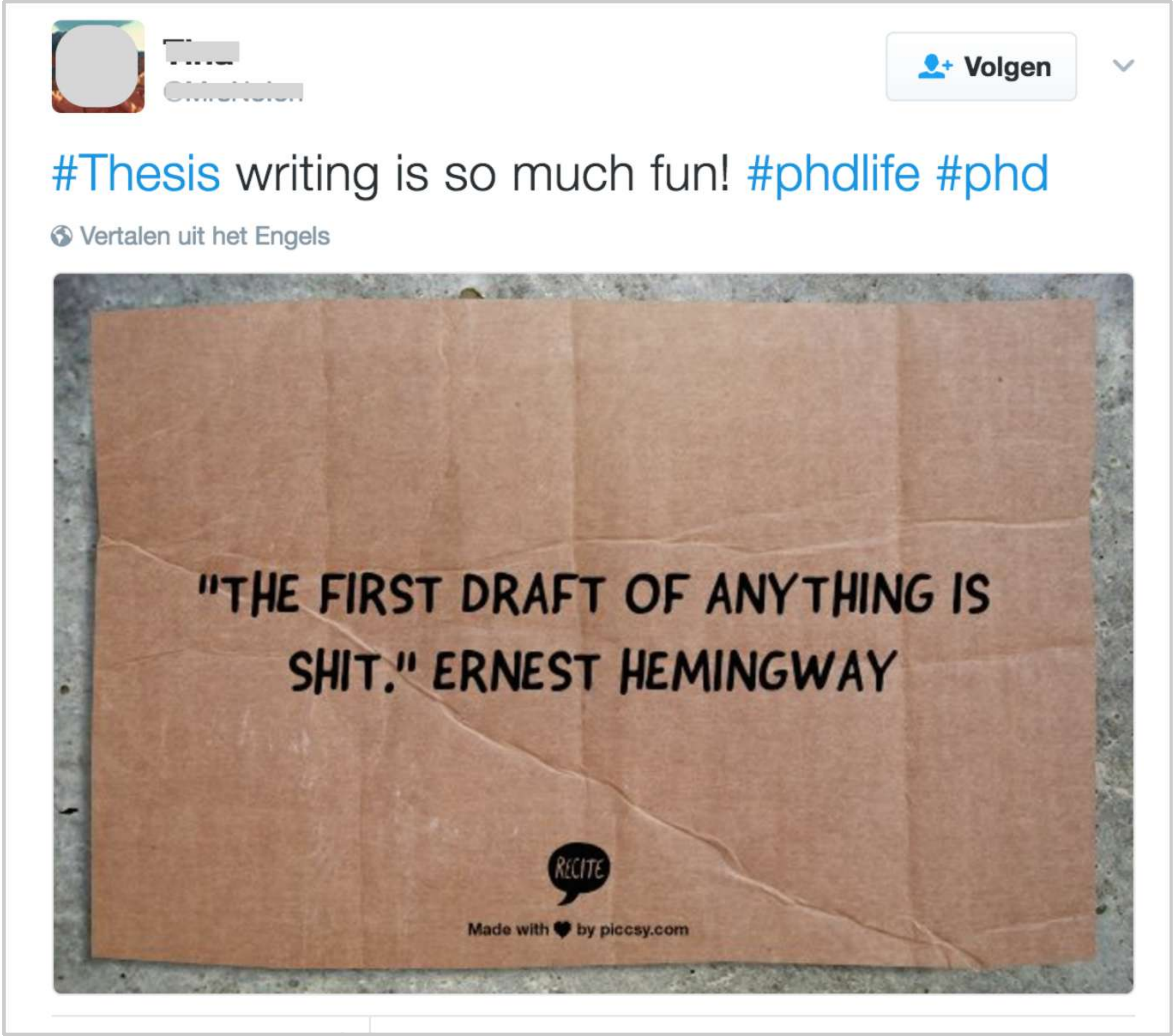
BUT:

- Ambiguity
  “Hi bitches, anyone in for a movie tonight?”
  “Shut up, you bitch!”
- Implicit realizations of cyberbullying
  “You make my fists itch…”
- Data sparseness

GHENT UNIVERSITY

IRONY DETECTION

GHENT UNIVERSITY



PhD of Cynthia Van Hee: ”Can machines sense irony?” (2017)

GHENT UNIVERSITY

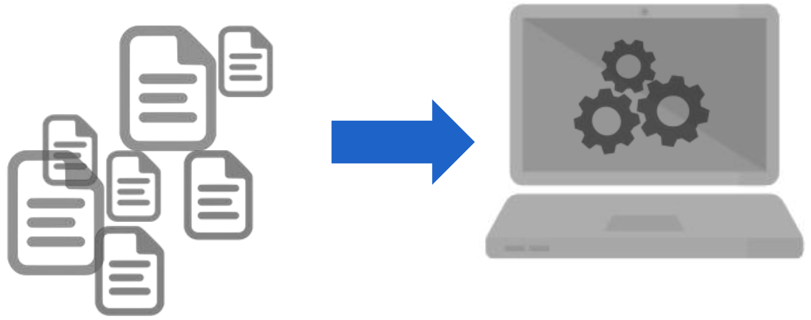
Corpus construction and annotation

Experiments:

- Exp. 1: Automatic irony detection
- Exp. 2: Modelling prototypical sentiment
- Exp. 3: Irony detection for sentiment analysis

CORPUS CONSTRUCTION

- Irony examples necessary to train the classifier



- Genre = Twitter
- Irony-related hashtags: *#not, #sarcasm, #irony*
- 3,000 English tweets (Van Hee et al., 2016a)

GHENT UNIVERSITY

Reference: Van Hee, C., Lefever, E. and Hoste, V.: 2016a, Exploring the realization of irony in Twitter data, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, pp. 1795–1799.

CORPUS ANNOTATION

- Manual annotations by trained linguists
- Task: which tweets are ironic and how is the irony realised?

literal sentiment: positive (“can’t wait”)

intended sentiment: negative (“go to the dentist”)

GHENT UNIVERSITY

ANNOTATION SCHEME

Fine-grained irony categories (Van Hee et al., 2016b)

1) Ironic by a polarity contrast

2) Situational irony

3) Other forms of verbal irony

4) Not ironic

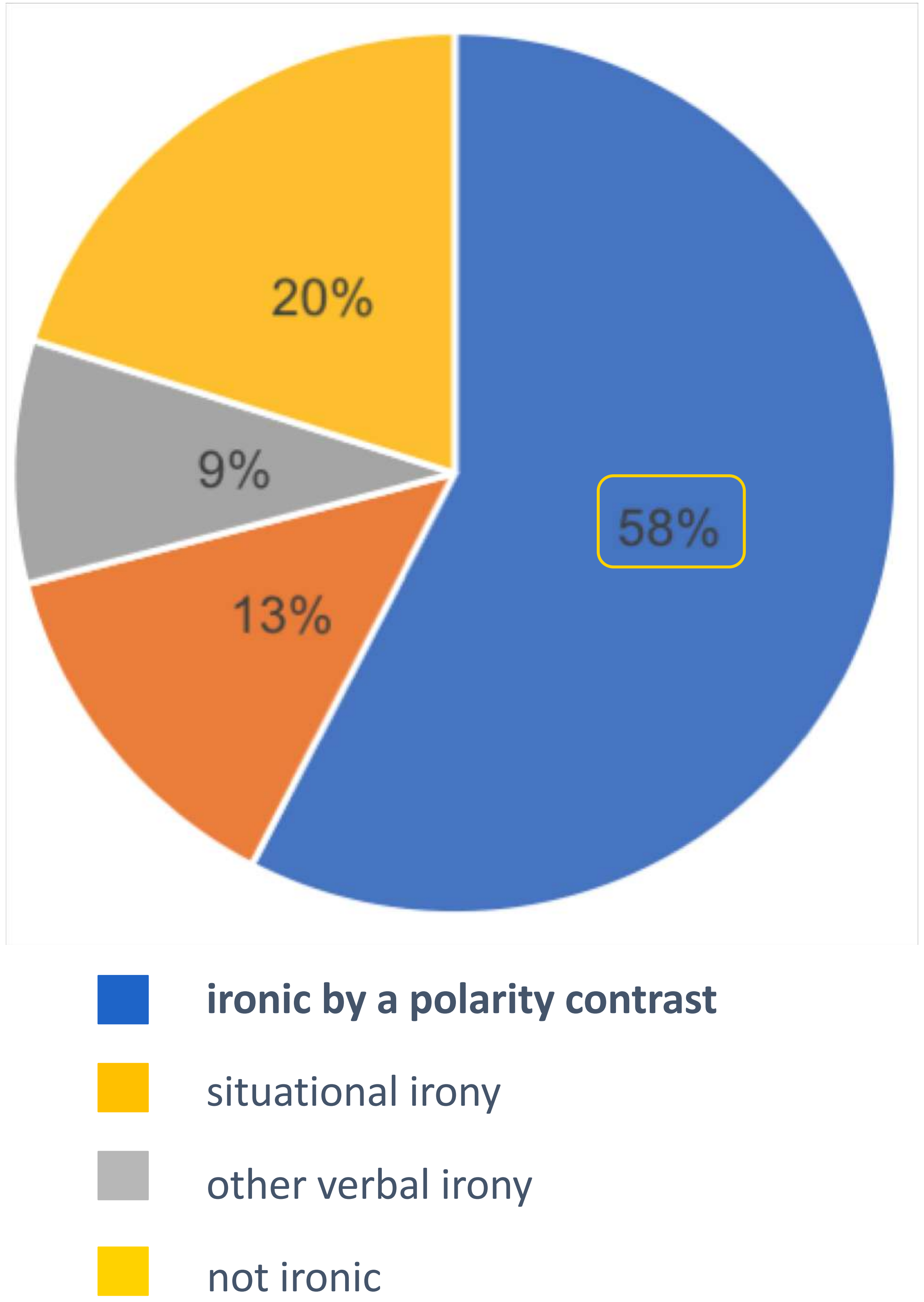
GHENT UNIVERSITY

Reference: Van Hee, C., Lefever, E. and Hoste, V.: 2016b, Guidelines for Annotating Irony in Social Media Text, version 2.0, *Technical Report 16-01*, LT3, Language and Translation Technology Team – Ghent University.

ANNOTATION SCHEME

- Fine-grained irony categories

1. Ironic by a polarity contrast
2. Situational irony
3. Other forms of verbal irony
4. Not ironic



GHENT UNIVERSITY

EXP. 1 AUTOMATIC IRONY DETECTION: HOW?

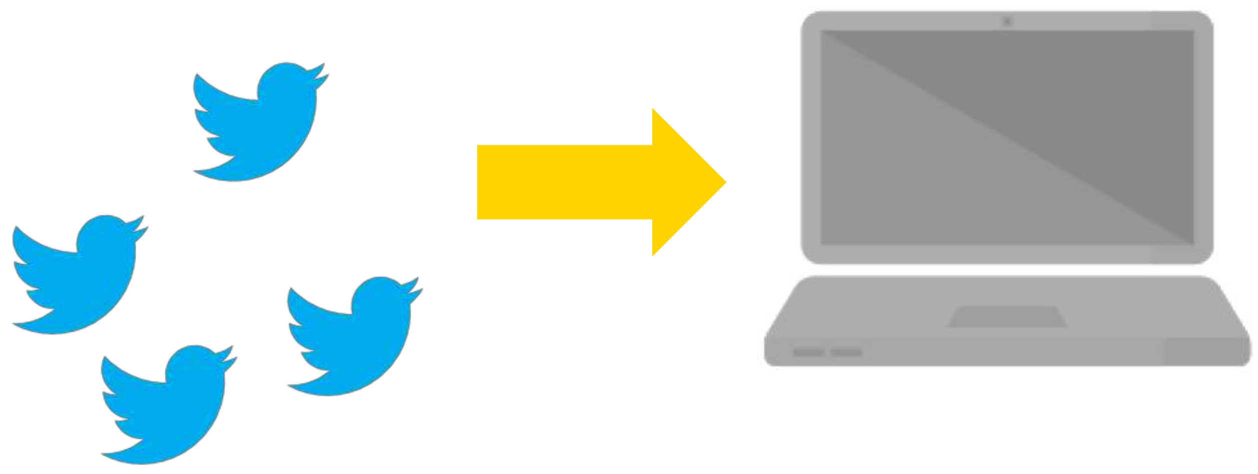
- Experimental corpus

3,000 tweets annotated corpus + extra non-ironic tweets for balanced distribution

- Preprocessing

Removal of hashtags ***#irony***, ***#not***, ***#sarcasm***

Supervised machine learning



GHENT UNIVERSITY

EXP. 1 AUTOMATIC IRONY DETECTION: FEATURES

LEXICAL: word & character sequences, character & punctuation repetition, emoticon frequency,…

[i love] - [love maths] - [lol] - [yea] - [yaaaaaay] - [??!!] - :-)

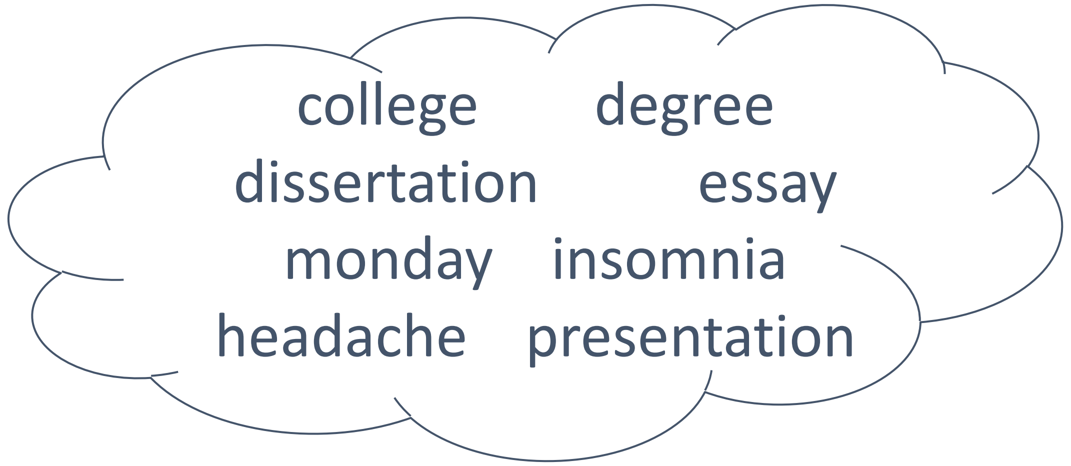
SYNTACTIC: part of speech frequencies, verb tenses, named entity frequencies

[V, A, N, #, E] - [past/present] - [people/location/organisation]

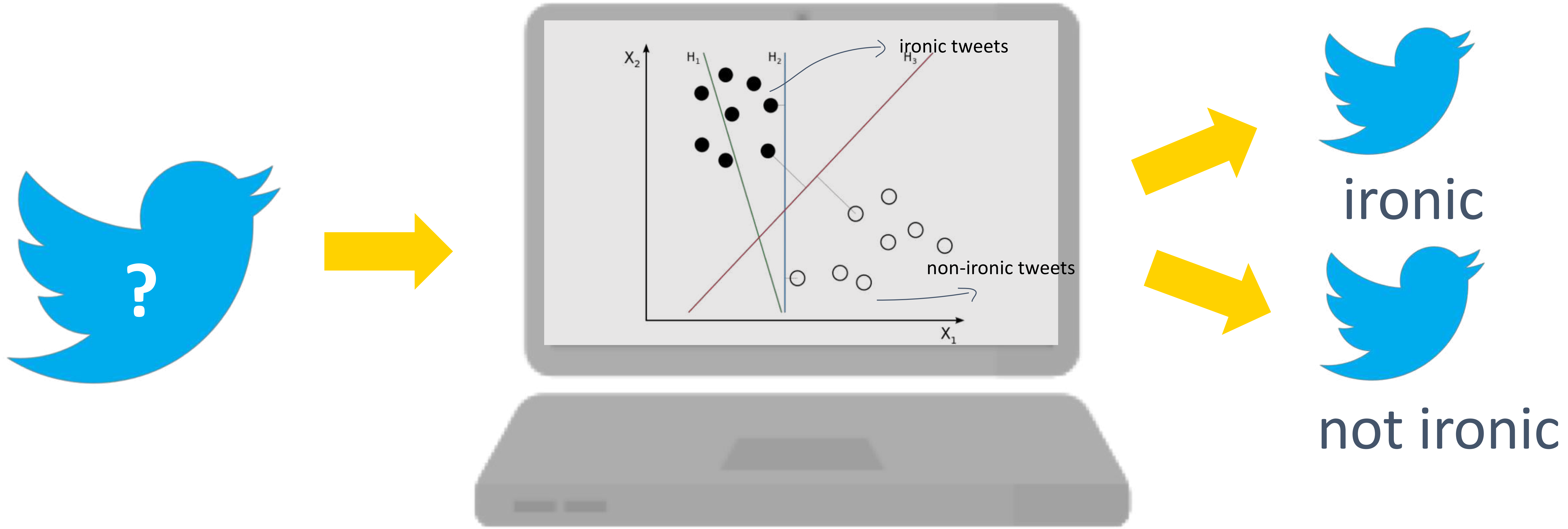
SENTIMENT: number of explicit positive/negative words

[hate] - [joyful] - [don’t like] - [bright]

SEMANTIC: semantic word clusters/topics



EXP. 1 AUTOMATIC IRONY DETECTION: HOW?



feature group	accuracy	precision	recall	F_1
lexical	**66.81**	**67.43**	66.60	**67.01**
sentiment	58.77	61.54	49.48	54.86
semantic	63.05	63.67	62.89	63.28
syntactic	64.82	64.18	**69.07**	66.53

GHENT UNIVERSITY

EXP. 1 AUTOMATIC IRONY DETECTION: BOTTLENECKS

- Tweets that carry implicit or **prototypical** sentiment

Just love to go to the dentist

explicitly positive

implicitly negative

EXP. 2 MODELLING PROTOTYPICAL SENTIMENT:

Input:

going to the dentist

not being able to sleep

car decides not to start

two hour flight delay

eight hour car ride

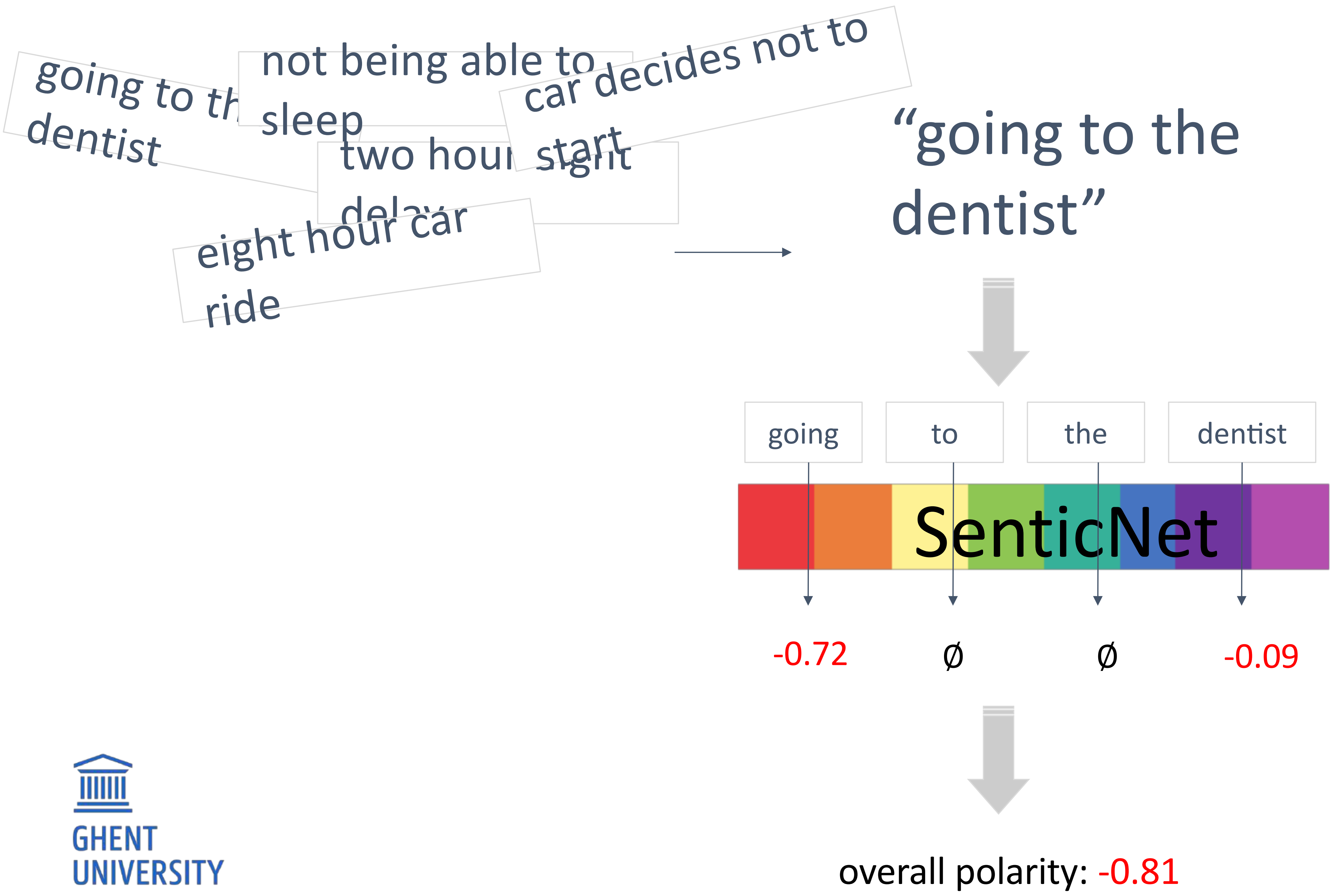
2 approaches:

SenticNet 4: lexical and semantics database (Cambria et al., 2016)

Twitter: resource of opinions shared in real time

GHENT UNIVERSITY

EXP. 2 MODELLING PROTOTYPICAL SENTIMENT: SENTICNET



GHENT UNIVERSITY

EXP. 2 MODELLING PROTOTYPICAL SENTIMENT: SENTICNET

Accuracy: **37%**

- \+ Fast and simple approach
- \- Focus on single words
- \- Rapidly evolving world → will coverage ever be sufficient?

EXP. 2 MODELLING PROTOTYPICAL SENTIMENT: TWITTER

"going to the dentist"

preprocessing

On my agenda tomorrow: going to the dentist

RT @someuser: uhu, I have to go take my wisdom teeth out #going to the dentist

"Yay weekend!", but NOOO, this gurl is going to the dentist first -_-

automatic sentiment analysis

polarity: negative

GHENT UNIVERSITY

EXP. 2 MODELLING PROTOTYPICAL SENTIMENT: TWITTER

Accuracy: **72%**

+ Look-up of multi-word phrases possible
+ Sentiment based on real-time ‘public’ opinion
- Sentiment based on real-time ‘public’ opinion
- Requires a large set of relevant tweets + automatic sentiment analysis system

Effect of crises, trends?

GHENT UNIVERSITY

EXP. 2 IRONY DETECTION: POLARITY CONTRAST APPROACH

- Lexical, syntactic, semantic + polarity contrast information
- Results: **improves irony detection performance**

	system	positive class	implicit sentiment	accuracy	precision	recall	F_1
	baseline SVM (lex+sem+synt)	ironic by clash + situational + other	-	**69.21%**	68.92%	71.34%	**70.11**
1	AND-combination	ironic by clash + situational + other	gold-standard	63.78%	**73.96%**	43.92%	55.11%
2	OR-combination	ironic by clash + situational + other	gold-standard	62.42%	59.15%	**83.30%**	69.18%
3	AND-combination	ironic by clash + situational + other	automatic	58.98%	69.01%	34.43%	45.94%
4	OR-combination	ironic by clash + situational + other	automatic	62.11%	58.97%	82.68%	68.84%

Precision: 69% → 74%

Recall: 71% → 83%

- Challenge: tweets that need more context: *“Excellent presentation #not”*

GHENT UNIVERSITY

IRONY DETECTION FOR SENTIMENT ANALYSIS

- State-of-the-art sentiment analysis systems work well: F_1 = 68% (Rosenthal et al., 2017)
- Bottleneck: irony

"automatically defining whether a given piece of text is positive, negative or neutral"

GHENT UNIVERSITY

EXP. 3 IRONY DETECTION FOR SENTIMENT ANALYSIS

- Sentiment classifier exploiting a rich feature set (Van Hee et al., 2014)
- Ranked 16th among 50 submissions in SemEval-2014 (Rosenthal et al., 2014)
- Results:

Without irony detection

SMS2013 2.093 inst.	TWE2013 3.813 inst.	TWE2014 1.853 inst.	TWE2014Sarcasm 76 inst.	LiveJour.2014 1.142 inst.	full test 8.987 inst.
70.53%	66.36%	64.83%	16.58%	68.00%	67.28%

Reference: Van Hee, C., Van de Kauter, M., De Clercq, O., Lefever, E. and Hoste, V.: 2014, LT3: Sentiment Classification in User-Generated Content Using a Rich Feature Set, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14), Association for Computational Linguistics, Dublin, Ireland, pp. 406–410.

GHENT UNIVERSITY

EXP. 3 IRONY DETECTION FOR SENTIMENT ANALYSIS: RESULTS

- Sentiment classifier optimisation: system ranks **1st**
- Adding irony information to sentiment classifier
- Results:

Without irony detection

dataset	accuracy	precision	recall	F_1
Twitter2014Sarcasm	17.11%	30.76%	54.35%	16.58%

With irony detection

+42 % **+20 %**

dataset	accuracy	precision	recall	F_1
Twitter2014Sarcasm	59.21%	40.56%	69.81%	36.71%

GHENT UNIVERSITY

SHORT OVERVIEW OTHER USE CASES

GHENT UNIVERSITY

ASPECT-BASED SENTIMENT ANALYSIS

GHENT UNIVERSITY

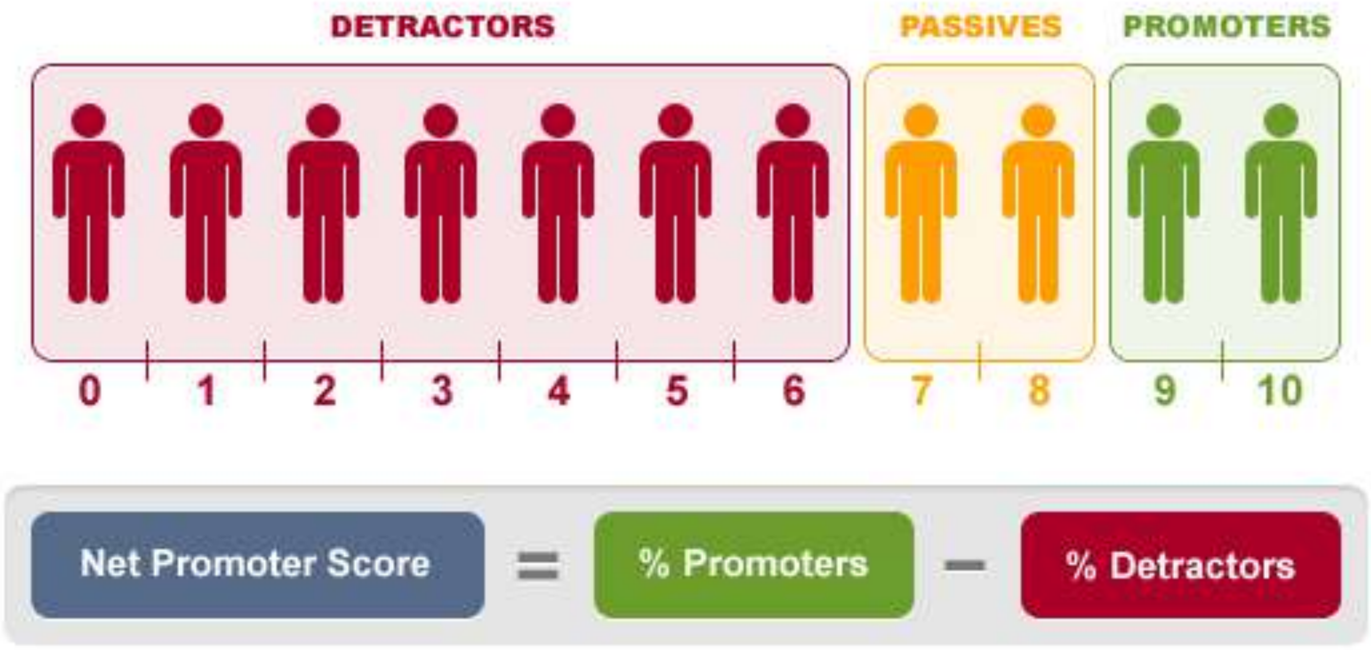
ASPECT-BASED SENTIMENT ANALYSIS

COLLECT DIRECT CUSTOMER FEEDBACK

“On a scale of 0 to 10, how would you recommend X to a friend or family?”

➡ QUANTITATIVE DATA

Trademark of Bain & Company, Inc and Fred Reichfeld



GHENT UNIVERSITY

ABSA

2. “Why did you assign this score?”

➡ QUALITATIVE DATA

FREE TEXT

» Writing in their own language (various languages, dialect, typos, …)

» Expressing **sentiments** about a variety of **aspects**

Store: PERSONNEL, COLLECTION, COMMUNICATION, …

GHENT UNIVERSITY

ABSA

ASPECT-BASED SENTIMENT ANALYSIS (ABSA)

i. Extract all aspect expressions of the entities
ii. Categorize these aspect expression into predefined categories
iii. Determine whether an opinion on an aspect is positive, negative or neutral

➡ SUPERVISED MACHINE LEARNING

ABSA

WHAT are they talking about?

HOW are they talking about it?

INTEGRATED PIPELINE WITH GOOD RESULTS

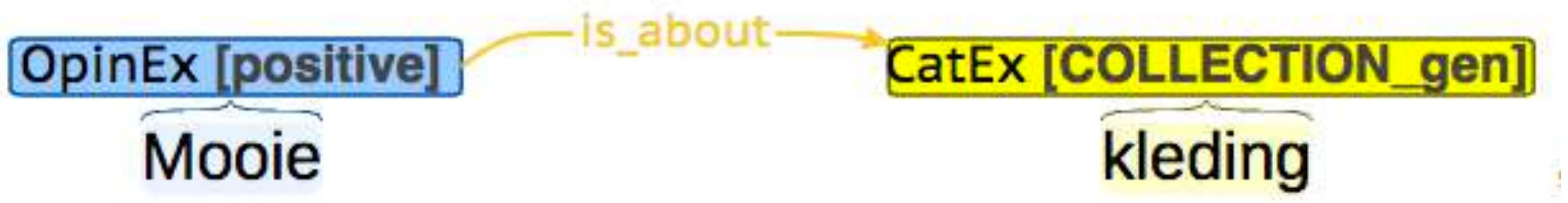
WHAT? → ASPECT TERM EXTRACTION (which words?)
→ ASPECT CATEGORY CLASSIFICATION (which aspects?)

HOW? → ASPECT POLARITY CLASSIFICATION (sentiment?)

GHENT UNIVERSITY

DOMAINS: BANKING, RETAIL, HR

DATASETS AND ANNOTATIONS: banking, retail, HR



e.g. RETAIL:
24 aspects

MARCOM	Communication, promotion
PERSONNEL	Advice, availability, expertise, reception, service
PRODUCT	Price, variety, kids, general, men, women, colour, sizes, quality, fit, pricequality
STORE	Fitting rooms, parking, general
CUSTSERVICE	General
BRANDS	General
WEBSHOP	General

GHENT UNIVERSITY

INTEGRATED ABSA PIPELINE

1. ASPECT TERM EXTRACTION

Sequential IOB labeling task

- Token shape features (capitalization, digits, alphanum, suffix)
- Lemma, PoS, chunk and NE label (LeTs preprocess, Van de Kauter et al. 2014)
- CRF Suite (LBFGS optimization function): 90% train - 10% test

INTEGRATED ABSA PIPELINE

2. ASPECT CATEGORY CLASSIFICATION

Multiclass classification

- Lexical features: bag-of-words (token unigrams)
- Lexical-semantic features: Dutch WordNet (Cornetto, Vossen et al. 2013) & DBPedia (Lehmann et al. 2013)
- LibSVM: 90% train - 10% test
- Output from previous step used as input for this step

INTEGRATED ABSA PIPELINE

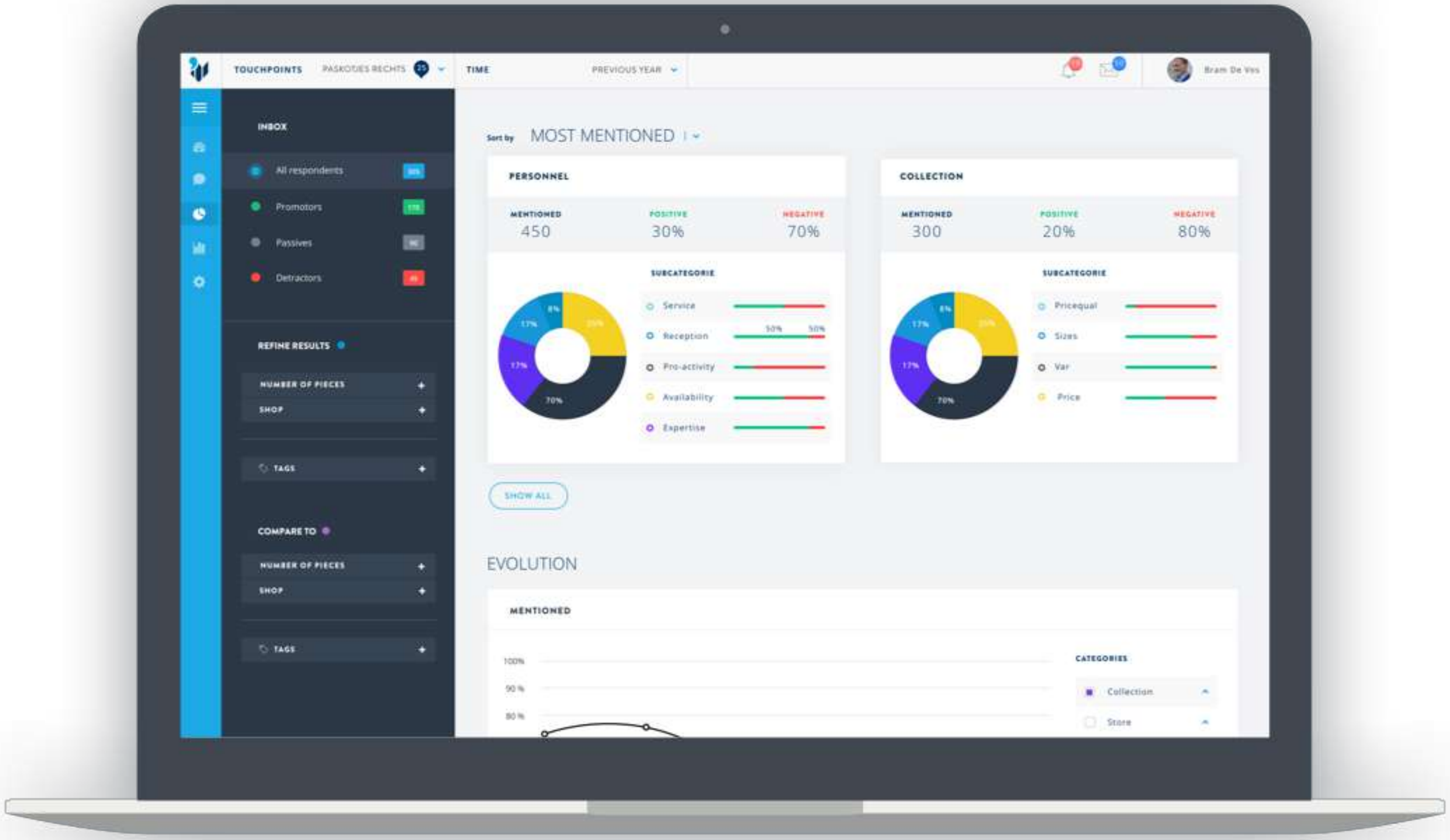
3. ASPECT POLARITY CLASSIFICATION

Multiclass classification (☺ 😐 ☹)

- Bag-of-words (token unigrams), predicted aspect category
- Lexicon-lookup features: training, Pattern (De Smedt and Daelemans 2012) & DUOMAN (Jijkoun and Hoffman 2009) + NEGATION
- LibSVM: 90% train - 10% test
- Output from previous two steps used as input for this step

GHENT UNIVERSITY

COOPERATION WITH HELLO CUSTOMER

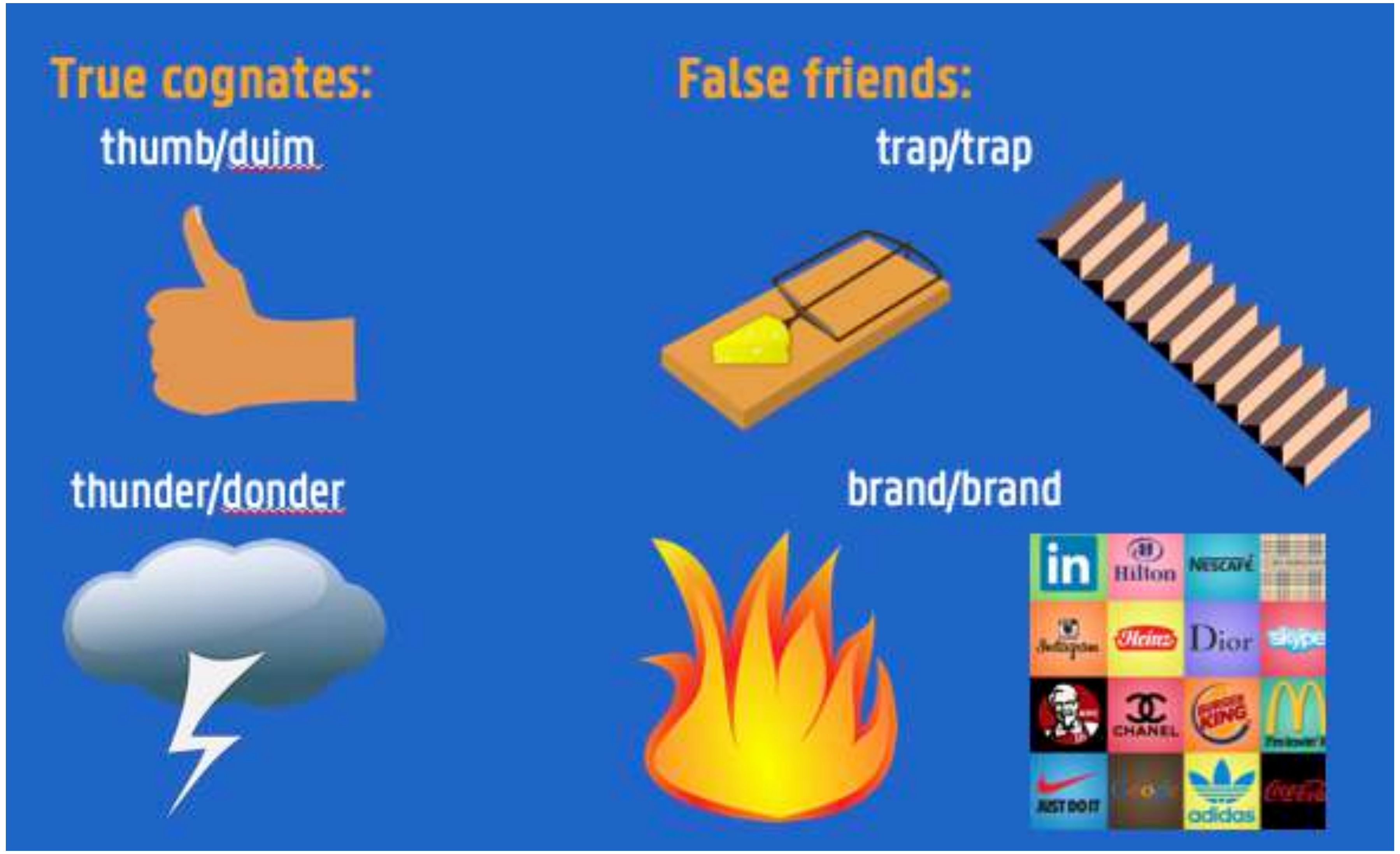


GHENT UNIVERSITY

Reference: De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., & Hoste, V. (2017). Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 136–142).

COGNATE DETECTION

COGNATES /VS/ FALSE FRIENDS



Cognates:
words with high formal and semantic cross-lingual similarity

False friends:
words which have similar forms, but differ in their meaning

GHENT UNIVERSITY

COGNATE DETECTION

Cognate detection = task to distinguish cognates from non-cognates (non-related words + false friends)

Use:

- Important skill for second language learners (CALL)
- Boost the performance of automatic alignment between related languages
- Compile bilingual lexicons

GHENT UNIVERSITY

GOALS

New gold standard

- Context-independent
- English-Dutch, French-Dutch

Supervised binary classifier

- Perform cognate detection
- Orthographic & semantic similarity information
- Binary: no distinctions made for false friends and non-equivalents words

Reference: Lefever, E., Labat, S. And Singh P. (2020) “Identifying Cognates in English-Dutch and French-Dutch by means of Orthographic Information and Cross-lingual Word Embeddings”, LREC 2020.

1. ORTHOGRAPHIC SIMILARITY FEATURES

- 15 different string similarity metrics (Frunza and Inkpen 2007)
- Measure formal relatedness between source and target words
- Metrics:

  Prefix, Dice (4 variants), Longest Common Subsequence Ratio, Normalized Levenshtein Similarity, Jaccard index, Jaro-Winkler similarity, Spsim (learns grapheme mappings between language pairs, Gomes and Pereira Lopes, 2011)

GHENT UNIVERSITY

2. Semantic information

Measure the semantic similarity between word pairs

Embeddings

- Pre-trained fastText embeddings (Common Crawl & Wikipedia)
- **Incremental re-training** (Grave et al., 2018) with domain-specific information
  → Accommodate for unseen words
- **Unsupervised mapping** in common vector space (Artetxe et al., 2018)
  → transformation matrix initialized by Singular Value Decomposition
  → train iteratively
- Cosine similarity

GHENT UNIVERSITY

RESULTS

	Cognate			Non-cognate			Average score		
Experiment	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
Ortho	0.909	0.992	0.952	0.909	0.798	0.850	0.909	0.895	0.902
Sem	0.997	1.000	0.998	0.987	0.422	0.672	0.997	0.711	0.830
Ortho + sem	0.915	0.993	0.955	0.915	0.793	0.853	0.915	0.893	0.904

Table 1: Precision (Prec), Recall (Rec) and F1-score for the classifier trained on English-Dutch data

	Cognate			Non-cognate			Average score		
Experiment	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
Ortho	0.951	0.940	0.945	0.929	0.810	0.864	0.940	0.875	0.905
Sem	0.915	1.000	0.956	0.925	0.642	0.764	0.920	0.821	0.868
Ortho + sem	0.943	1.000	0.971	0.943	0.804	0.879	0.943	0.908	0.925

Table 2: Precision (Prec), Recall (Rec) and F1-score for the classifier trained on French-Dutch data

GHENT UNIVERSITY

ANALYSIS

Semantic information helps to:

- detect cognate pairs showing less orthographic resemblance (*older-ouderen, widespread-wijdverbreid, sweating-zweten, shame-schaamte*)
- generate less false negatives. Wrongly labeled by the classifier relying on orthographic information: *affects-effecten, unlocking-blokkering, provide-profielen, where-wateren*
- Generate few additional false negatives (*include-inhouden, docker-dokwerker*) and false positives (*told-toen, because-bepaalde*)

GHENT UNIVERSITY

WINE CLASSIFICATION

EXPERTS WRITING WINE REVIEWS

Wine experts convert sensory input to words on a daily basis

Cantina del Pino makes some of the finest Barbaresco available today. This shows a succulent quality, with aromas of smoked bacon, wild berries and forest underbrush. Savory and sophisticated, this has loads of personality.

Red, 2009, Nebbiolo grape, price $45, Italy, rating of 91

GHENT UNIVERSITY

RESEARCH QUESTIONS

1. Do wine experts share a common vocabulary, or is it just "purple prose" (Quandt, 2007)?
   – **What is the usefulness of domain-specific terminology as feature representations?**

⇒Classification experiments

GHENT UNIVERSITY

RESEARCH QUESTIONS

2. Is there a **correlation** between prices and (subjective) ratings of wines? Between ratings and review text? More expensive wines > more “expensive” (longer) words?

=> Regression analysis

CORPUS OF WINE REVIEWS

- from http://www.winemag.com
- Corpus of 76,410 unique reviews from 33 experts
- labeled with meta data (price, color, producer, grape, etc)
- short reviews (39 words on average)
- rating between 80 and 100
- we only use the reviews without missing values

GHENT UNIVERSITY

CLASSIFICATION EXPERIMENTS

- Goal: automatically predict objective wine characteristics: color, grape variety, and country of origin
- Experimental set-up:
  - Supervised machine learning: SVM
  - Train (80%) – Test (20%)

Classification Task	Training	Test	Categories
Color	56,893	14,209	3
Grape type	39,900	9,976	28
country	61,128	15,282	47

GHENT UNIVERSITY

INFORMATION SOURCES

- Linguistically preprocessed (Stanford toolkit)
- Three different feature types:
  1. Lexical features (BoW)
  2. Semantic features (word embeddings)
  3. Terminology features (TExSIS)

GHENT UNIVERSITY

LEXICAL FEATURES

- Bag-of-words unigram features
- Lowercased lemmas
- Filtered on PoS-tag (nouns, adjectives, verbs, adverbs)
- Incorporated as binary features

GHENT UNIVERSITY

SEMANTIC FEATURES

- **Word embeddings** from the training reviews (Word2Vec, Mikolov 2013): BoW model, context size=8, 200 features
- **Clustered** the resulting word vectors (group words that share common contexts in the wine reviews) using K-means clustering (300 clusters)
- Implemented resulting clusters as **binary features**

SEMANTIC FEATURES

⇒ Clusters indeed semantically related terms

⇒ E.g. cluster 82 (terms related to floral and other related aromas):

abundant, acacia, aromatic, bee's, clover, dandelion, delicate, enticing, floral, flower, foremost, fragrant, freesia, fresh-cut, freshly, fuzz, garden, jasmine, light-weight, lilac, musk, oils, peony, petroleum, pretty, roses, rosewater, subtle, talcum, wax, wisp, wispy

TERMINOLOGY FEATURES

- Wine-specific terms were extracted with TExSIS (Macken et al. 2013)

- hybrid term extractor:
  - Linguistic **preprocessing** (LeTs Preprocess, Van de Kauter et al., 2013)
  - Linguistic information > generate syntactically valid **candidate terms**
  - **Statistical filtering** (termhood, log-likelihood, c-value), intuition: domain-specific terms have higher relative frequency in the wine corpus than in standard corpus (Web 1T 5-gram corpus)
  - Incorporated 15,000 terms with **highest termhood** values as binary features

GHENT UNIVERSITY

TOP-20 TERMINOLOGY FEATURES

Term	Termhood
flavor	1359.38
tannin	1018.32
aroma	997.62
wine	935.61
acidity	929.98
fruit	814.73
palate	792.01
finish	590.85
off-dry	587.43
cherry	580.13

Term	Termhood
single-vineyard	503.82
tannic	489.55
mouthfeel	488.72
cool-climate	448.87
black-fruit	432.85
crisp	418.22
Port-like	237.21
tangy	210.88
crisp acidity	207.21
cherry fruit	185.69

CLASSIFICATION RESULTS

Setup	RBF opt	Lin Kernel
	Color	
BoW	96.75%	96.59%
Word2Vec	96.31%	96.18%
TExSIS	93.49%	91.66%
All features	96.09%	95.29%
	Grape Variety	
BoW	42.10%	48.28%
Word2Vec	57.39%	56.46%
TExSIS	72.53%	72.77%
All features	76.16%	76.61%
	Country	
BoW	66.50%	60.32%
Word2Vec	69.17%	68.27%
TExSIS	78.67%	79.06%
All features	82.27%	82.84%

CONCLUSION

– Wine experts indeed share a common vocabulary, making it possible to predict color, grape variety and country of origin

– Terminological features express sensory information

– Terminological features outperform BoW-features and semantic features

– Review length and avg. word length are significantly related to review price and rating

GHENT UNIVERSITY

DIGITAL HUMANITIES APPLICATIONS

GHENT UNIVERSITY

NLP FOR DIGITAL HUMANITIES

- Historical sentiment analysis
- Detect orthographic and semantic similarity between epigrams in medieval Greek

Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)

Language technology for digital humanities: introduction to the special issue

Erhard Hinrichs, Marie Hinrichs, Sandra Kübler & Thorsten Trippel

Language Resources and Evaluation **53**, 559–563(2019) | Cite this article

GHENT UNIVERSITY

FACULTY OF ARTS
AND PHILOSOPHY

Els Lefever

LT^3 Language and translation technology team

Ghent University

Interested in collaboration?

els.lefever@ugent.be

GHENT
UNIVERSITY