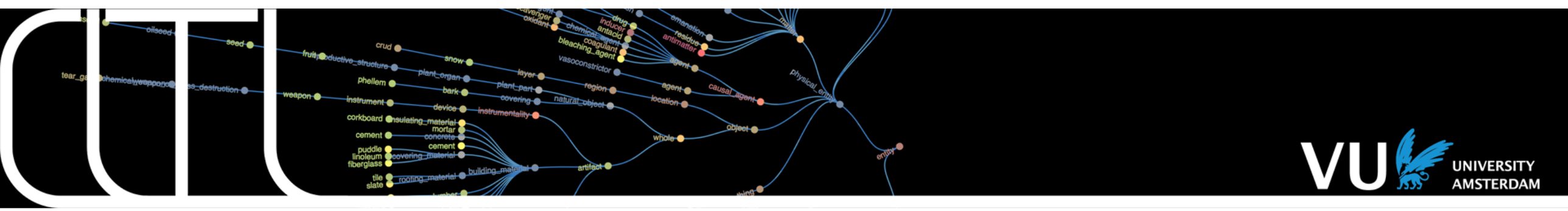
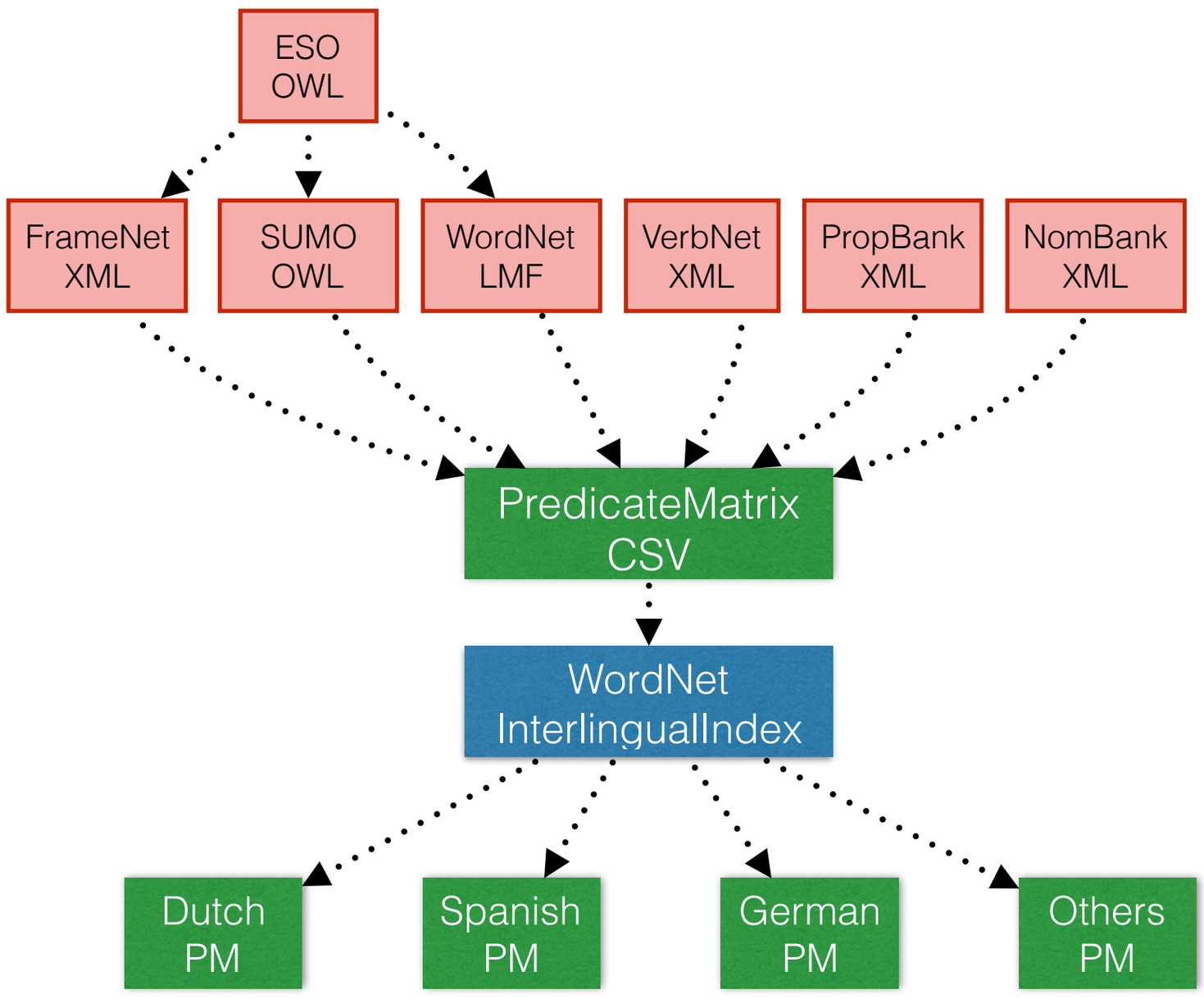
Open Dutch FrameNet

Piek Vossen
Computational Lexicology and Terminology Lab (CLTL)
VU University of Amsterdam

“Builders of the Open Dutch Wordnet and DutchSemCor”



Linking Lexical resources for event extraction from text



English SRL module analysis

Qatar Holding sells 10% stake in Porsche to founding families

Predicate Matrix

A0	sell.01	A1	A2	PropBank
A0	sale.01	A1	A2	NomBank
Buyer	Commerce_sell	Goods	Seller	FrameNet
Agent	give-13.1-1	Theme	Recipient	VerbNet
Possession-owner_1	Selling		Possession_owner_2	ESO
	ili-30-02242464-v			WordNet
arg0	venta.01	arg1	arg2	AnCora-Nom
arg0	vender.01	arg1	arg2	AnCora-Verb

El holding de Qatar vende su participación del 10% en Porsche a las familias fundadoras

Spanish SRL module analysis

Dutch PredicateMatrix

- 66376 Dutch lexical units mapped to SemLink data and others
- lu-lemma:**deponeren** odwn-eq_synonym:r_v-2182 vn-class:put_spatial-9.2 vn-class-nr:9.2 vn-lemma:lay wn:lay%2:35:01 **mcr:ili-30-01494310-v fn:Placing fn-entry:lay.v** pb-sense:lay.01 mcr-sumo:contact wn-file:24 wn-sense-nr:107 eso:Placing fn-pb-role:Agent#0 fn-pb-role:Cause#0 fn-pb-role:Theme#1 fn-pb-role:Goal#2

Why annotate if you have a Dutch PredicateMatrix?

- SemLink has low coverage
- PredicateMatrix has low precision
- Links between Dutch wordnet and English wordnet are not perfect (recall & precision)
- We cannot train nor test a proper FrameNet SRL system without annotating actual sentences

Overall strategy

- Full text strategy: all content words - so far only verbs - annotated (no preselection of lexical units; no preselection of example sentences):
    - annotation represent the corpus and not the lexicon
        - identify the frame for a given predicate that fits the sentence, and assign the corresponding roles.
- Annotators were provided with detailed guidelines
- We developed our own annotation tool with access to parsed text and to the English FrameNet through Dutch and English words

Overall strategy

- All frames strategy (no pre-selection of frames): the annotator chooses the correct frame from one of the 1300+ frames from the English FN (version 1.7)

- Annotations are added on top of PropBank annotations (events and arguments are already identified).

- no new frames are added; if a frame could not be found the event is annotated with ’none’ all files are double-annotated

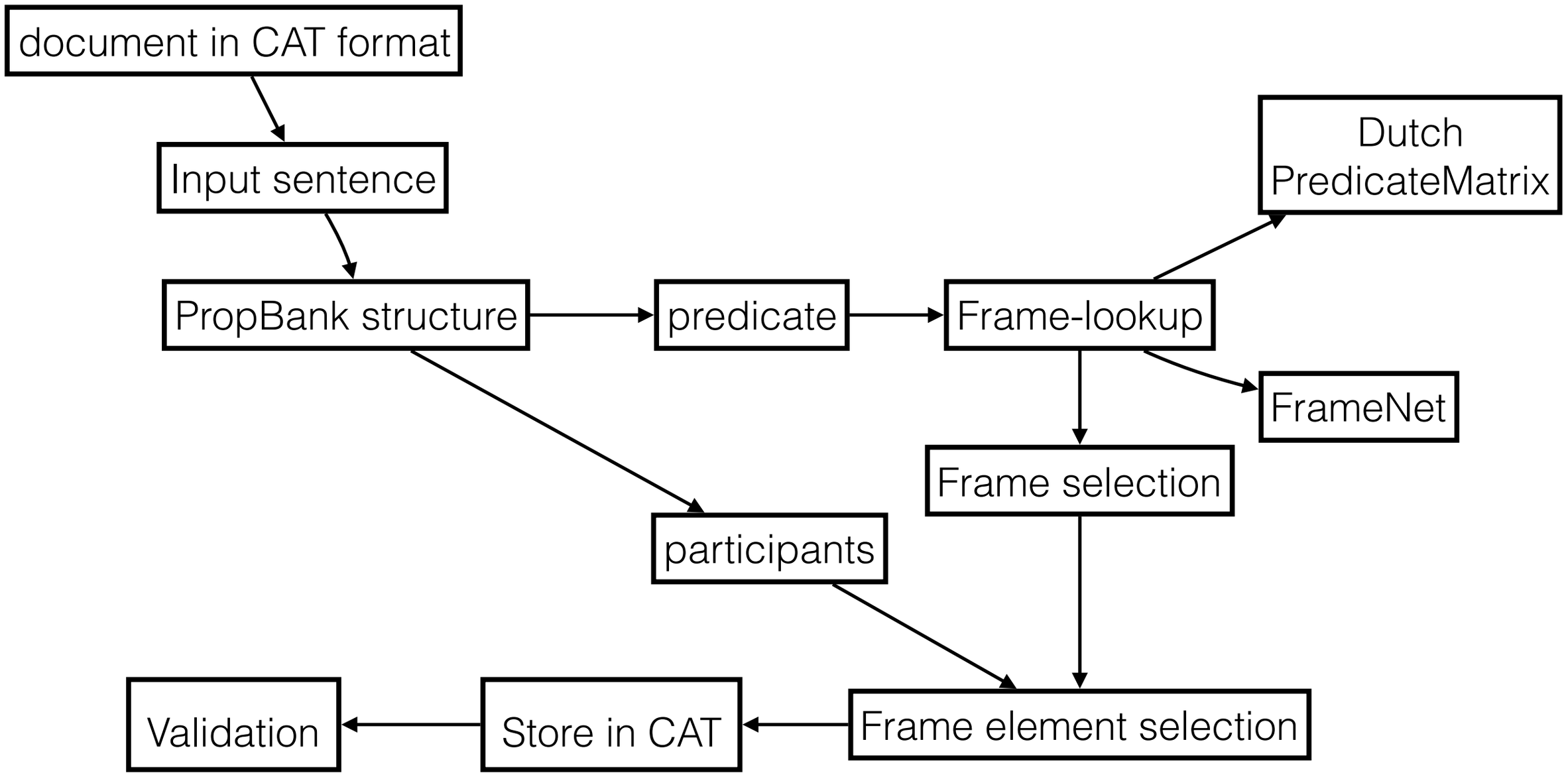
- Special cases: metaphors and multiwords

- Validation of disagreements using the CAT tool, where we mark all disagreeing annotations and related annotations

Annotation tool

- https://github.com/cltl/FrameNet-annotation-tool
- Uses:
    - FrameNet lexical unit index and Frames
        - PredicateMatrix (derived from SemLink,de Lacalle et al 2014) mapped to Dutch lexical units through wordnet links
- Assumes input files in CAT format with Propbank annotations marking predicates, roles and role labels
- Command-line interface, iterates over sentences in documents stored locally

Annotation process



Annotation tool

```
------------------------- EXPLANATION --------------------------

There are several options:
(1) Enter the number of the correct frame element.
(2) Enter multiple numbers separated by commas if you want to compare some definitions first.
(3) Enter None if none of the roles is the correct one.
(4) Enter WrongRelation if there is something wrong with this particular relation (e.g. this is not an argument of this predicate)
.


---------------------- ANNOTATION OF ROLE ----------------------

SENTENCE: De vier buitenplaneten stonden toen op een lijn .
PREDICATE: stonden
ARGUMENT: De vier buitenplaneten

------------------------------------------------------------

YOU HAVE CHOSEN:  Being_located

THE POSSIBLE ROLES FOR THIS FRAME ARE:
0 Theme
1 Place
2 Dependent_state
3 Time
4 Location
5 Cotheme
6 Depictive

PLEASE ENTER THE NUMBER(S) OF THE ROLE OF THE ARGUMENT: 0
```

File | Task | Markable | Relation | Statistics | Help — CATool

Corpus | Confirm Markable | Delete Selection | Task Selection: FN | HAS_PARTICIPANT

/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
/Users/antskez/SRLexam...
FB-FactMarkers
FB-microprc
FN-48-Vince
FN-48b-casp
Framenet
FrameNet2
Joran
joran-works
LREC
LREC-2
MASC
MicroPropos
MicroPropos
Round1-Uns
Round1-Uns
Round1-Uns
Round1-Uns
Round2-Uns
Round2-Uns

Markables
ENTITY_ME
EVENT_MEN

Relation Attributes

Relation type:	HAS_PARTICIPANT
frame:	Ingestion
sem_role:	ArgM-LOC
frame_element:	Place

*unsaved values

Save

S55 Engelsen en Belgen kennen de bovengisting (hoge gisting) , hetgeen **resulteert in verschillende biersoorten** .

S56 Bovengisting wordt tegenwoordig ook elders op het vasteland toegepast .

S57 Veel Belgische bieren zijn nog steeds bovengistend , en in Nederland bijvoorbeeld de bieren van de Amsterdamse Brouwerij 't IJ , die in Vlaanderen de mosterd haalde .

S58 De rond Brussel gesitueerde lambiekbrouwerijen nemen wat **het gistingsproces betreft** een aparte plaats in : doordat **men de wort een etmaal in open kuipen aan de buitenlucht blootstelt** , zorgen in de lucht aanwezige gisten (" wilde gisten ") en in eiken tonnen huizende micro-organismen voor een zogenaamde spontane vergisting .

S59 Geschiedenis

S60 Bier werd waarschijnlijk al 4500 jaar voor Christus gedronken in Mesopotamië .

wen werd bier nog altijd volgens dezelfde oude methode gebrouwen , voornamelijk door vrouwen , en in kloosters en abdijen .

dijen een zware en een lichte versie van het bier , waarbij de zware versie werd gedronken door paters en gasten , en de lichtere versie door de zusters .

t brouwen overgenomen door herbergen .

met kruidenmengsels , ook wel gruit genaamd .

Relations

S58 nemen <-> **HAS_PARTICIPANT** <-> **S58** De rond Brussel gesitueerde lambiekbrouwerijen
S58 nemen <-> **HAS_PARTICIPANT** <-> **S58** een aparte plaats
S58 nemen <-> **HAS_PARTICIPANT** <-> **S58** wat het gistingsproces betreft
S58 blootstelt <-> **HAS_PARTICIPANT** <-> **S58** men
S58 blootstelt <-> **HAS_PARTICIPANT** <-> **S58** een etmaal
S58 blootstelt <-> **HAS_PARTICIPANT** <-> **S58** in open kuipen
S58 blootstelt <-> **HAS_PARTICIPANT** <-> **S58** aan de buitenlucht
S58 blootstelt <-> **HAS_PARTICIPANT** <-> **S58** de wort
S58 zorgen <-> **HAS_PARTICIPANT** <-> **S58** in de lucht aanwezige gisten wilde gisten en in eiken tonnen huizende micro-organismen
S58 zorgen <-> **HAS_PARTICIPANT** <-> **S58** voor een zogenaamde spontane vergisting
S58 zorgen <-> **HAS_PARTICIPANT** <-> **S58** doordat men de wort een etmaal in open kuipen aan de buitenlucht blootstelt
S60 gedronken <-> **HAS_PARTICIPANT** <-> **S60** in Mesopotamië
S60 gedronken <-> **HAS_PARTICIPANT** <-> **S60** waarschijnlijk
S60 gedronken <-> **HAS_PARTICIPANT** <-> **S60** al 4500 jaar voor Christus
S61 gebrouwen <-> **HAS_PARTICIPANT** <-> **S61** volgens dezelfde oude methode
S61 gebrouwen <-> **HAS_PARTICIPANT** <-> **S61** voornamelijk door vrouwen en in kloosters en abdijen
S61 gebrouwen <-> **HAS_PARTICIPANT** <-> **S61** In de vroege Middeleeuwen

Project overview

- Preparations:
  - SoNaR corpus with manual PropBank annotations were developed in the national STEVIN project
  - Chantal Van Son developed the annotation tool as her Master thesis project
    - Dutch predicate matrix constructed in EU project NewsReader
- Personal fund to extend the PropBank annotation with FrameNet frames and elements:
  - 4 student assistants have been working for 6 months (8 hours/week)
  - PostDoc supervised the project
  - Annotation will end by May 2017

Corpus: SoNaR-klein 825K tokens

Table 1: Corpus statistics

theme/genre	nr_of_files	nr_of_annotated_verbs
background-news	2	51
financial	9	904
medical	1	88
news	5	499
newsletter	3	111
periodicals	37	821
policy	12	352
teletext	3	169
websites	1	49
wiki	28	854
totals	101	3898

Annotation statistics

Table 2: FN annotation statistics and agreement

nr of verbs (tokens)=nr of frames	3898	
nr of unique verbs (types)	1119	
nr of unique frames	651	
strict agreement on frames	47 %	$\kappa = 0.463$
lenient agreement on frames	55 %	
agreement frame elements (with matching frames)	79 %	

- strict agreement: only identical frames are considered matching (percentage agreement)
- lenient agreement: frames that are related through framenet relations 'inheritance' or 'see also' or 'uses' relations are considered matching.
- agreement on frame elements: measured only when frames are matching

Table 3: top 20 frame confusions

19	Activity_start	Process_start
14	Creating	Intentionally_create
14	Cause_change_of_position_on_a_scale	Change_position_on_a_scale
12	Using	Using_resource
12	Opinion	Regard
10	Cooking_creation	Manufacturing
8	Getting	Receiving
8	Expressing_publicly	Statement
8	Existence	Presence
8	Awareness	Grasp
7	Operate_vehicle	Self_motion
7	Finish_competition	Finish_game
7	Causation	Evidence
7	Being_named	Name_conferral
6	Perception_active	Perception_experience
6	Intentionally_create	Text_creation
6	Giving	Grant_permission
6	Cure	Medical_intervention
6	Cause_to_perceive	Expressing_publicly
6	Beat_opponent	Finish_competition
6	Awareness	Certainty
6	Accomplishment	Getting
5	Reference_text	WrongRelation
5	Preventing	Thwarting
5	Perception_active	Reference_text
5	Have_associated	Possession
5	Finish_competition	Success_or_failure
5	Competition	Finish_competition
5	Communication	Statement
5	Communication	Expressing_publicly

Table 4: agreement on frequent frames

frame	agreements	disagreements	percentage agreement
Statement	108	60	0.64
Possession	38	64	0.37
Existence	26	62	0.3
Causation	38	45	0.46
Event	30	46	0.39
Coming_to_be	30	46	0.39
Intentionally_act	24	49	0.33
Change_position_on_a_scale	27	39	0.41
Awareness	14	50	0.22
Activity_start	23	35	0.4
Being_located	22	35	0.39
Intentionally_create	4	52	0.07
Opinion	18	38	0.32
Receiving	22	31	0.42
Inclusion	12	39	0.24
Attempt_suasion	33	18	0.65
Evidence	14	34	0.29
Cause_to_perceive	11	36	0.23
Becoming_aware	12	33	0.27
Using	17	28	0.38
None	2	40	0.05
Giving	7	35	0.17
Perception_active	17	24	0.41
Finish_competition	5	35	0.12
Self_motion	17	23	0.42
Participation	12	27	0.31
Request	19	19	0.5
Communication	6	30	0.17
Desiring	25	11	0.69
Circumscribed_existence	2	32	0.06
Process_start	8	26	0.24
Arriving	8	26	0.24
Cause_change	13	21	0.38
Removing	16	18	0.47
Accomplishment	6	27	0.18

- Dutch verb lexicon with frames:
  - 3858 lexical units
  - 2348 entries

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<fnLexicon lang="nl">
<ENTRY lemma="inschakelen" pos="v">
        <frameAnnotation frame="Installing" annotations="1"/>
        <frameAnnotation frame="Process_start" annotations="1"/>
</ENTRY>
<ENTRY lemma="mankeren" pos="v">
        <frameAnnotation frame="Medical_conditions" annotations="2"/>
        <frameAnnotation frame="Undergoing" annotations="2"/>
</ENTRY>
<ENTRY lemma="baseren" pos="v">
        <frameAnnotation frame="Evidence" annotations="5"/>
        <frameAnnotation frame="Justifying" annotations="1"/>
        <frameAnnotation frame="None" annotations="1"/>
        <frameAnnotation frame="Reliance" annotations="3"/>
</ENTRY>
<ENTRY lemma="gemaakt/maken" pos="v">
        <frameAnnotation frame="Building" annotations="2"/>
        <frameAnnotation frame="Causation" annotations="1"/>
        <frameAnnotation frame="Cooking_creation" annotations="4"/>
        <frameAnnotation frame="Cotheme" annotations="1"/>
        <frameAnnotation frame="Create_representation" annotations="6"/>
        <frameAnnotation frame="Differentiation" annotations="2"/>
        <frameAnnotation frame="Fame" annotations="1"/>
        <frameAnnotation frame="Intentionally_create" annotations="1"/>
        <frameAnnotation frame="Manufacturing" annotations="2"/>
        <frameAnnotation frame="Physical_artworks" annotations="2"/>
        <frameAnnotation frame="Scrutiny" annotations="1"/>
        <frameAnnotation frame="Seeking" annotations="1"/>
        <frameAnnotation frame="Statement" annotations="2"/>
        <frameAnnotation frame="Success_or_failure" annotations="2"/>
        <frameAnnotation frame="Thwarting" annotations="1"/>
        <frameAnnotation frame="Undergo_change" annotations="1"/>
</ENTRY>
```

```
<ENTRY lemma="stellen" pos="v">
    <frameAnnotation frame="Activity_stop" annotations="2"/>
    <frameAnnotation frame="Appointing" annotations="1"/>
    <frameAnnotation frame="Attempt" annotations="2"/>
    <frameAnnotation frame="Cause_motion" annotations="1"/>
    <frameAnnotation frame="Cause_to_perceive" annotations="1"/>
    <frameAnnotation frame="Confronting_problem" annotations="1"/>
    <frameAnnotation frame="Employing" annotations="2"/>
    <frameAnnotation frame="Have_as_requirement" annotations="4"/>
    <frameAnnotation frame="Having_or_lacking_access" annotations="1"/>
    <frameAnnotation frame="Intentionally_act" annotations="1"/>
    <frameAnnotation frame="Intentionally_create" annotations="2"/>
    <frameAnnotation frame="MWE" annotations="2"/>
    <frameAnnotation frame="None" annotations="1"/>
    <frameAnnotation frame="Offering" annotations="1"/>
    <frameAnnotation frame="Opinion" annotations="1"/>
    <frameAnnotation frame="Possession" annotations="1"/>
    <frameAnnotation frame="Questioning" annotations="2"/>
    <frameAnnotation frame="Speak_on_topic" annotations="2"/>
    <frameAnnotation frame="Statement" annotations="5"/>
    <frameAnnotation frame="Supply" annotations="1"/>
    <frameAnnotation frame="Telling" annotations="2"/>
    <frameAnnotation frame="Text_creation" annotations="2"/>
    <frameAnnotation frame="Using" annotations="1"/>
    <frameAnnotation frame="Work" annotations="1"/>
</ENTRY>
```

Availability

- Our annotations are open source and freely available
- but the original texts can only be obtained through a license:
  - SoNaR-klein-commercieel enriched with PropBank annotations (and others)
  - Commercial license, free for research (0 euro)
- Wikipedia articles can be distributed freely
- http://tst-centrale.org/nl/tst-materialen/corpora/sonar-klein-corpus-commercieel-detail (Unfortunately only in Dutch so ask us for help)

Future plans

- Annotation of nominal events
- Apply for national funding (September 2017) to build a Dutch FrameNet resource and FrameNet automatic annotation tool
- Different approaches:
    - Resource linking: wordnets, ESO ontology, PredicateMatrix
    - Unsupervised learning with seed data: word embeddings and NN.
    - Lexicological checking and editing of the FrameNet lexicon
- Available in 2021 if we get the funding