Swedish multiword expressions and sublanguage parsing

DIMITRIOS KOKKINAKIS
Dep. of Swedish, the Swedish Language Bank &
the Centre for Language Technology (CLT)
University of Gothenburg
Sweden
dimitrios.kokkinakis@svenska.gu.se

OVERVIEW

- Types of MWE
- Textual data
- Related work
- Parsing method
- MWE knowledge integration
- Goal: event extraction
- Results and some future work

Plethora of MWEs

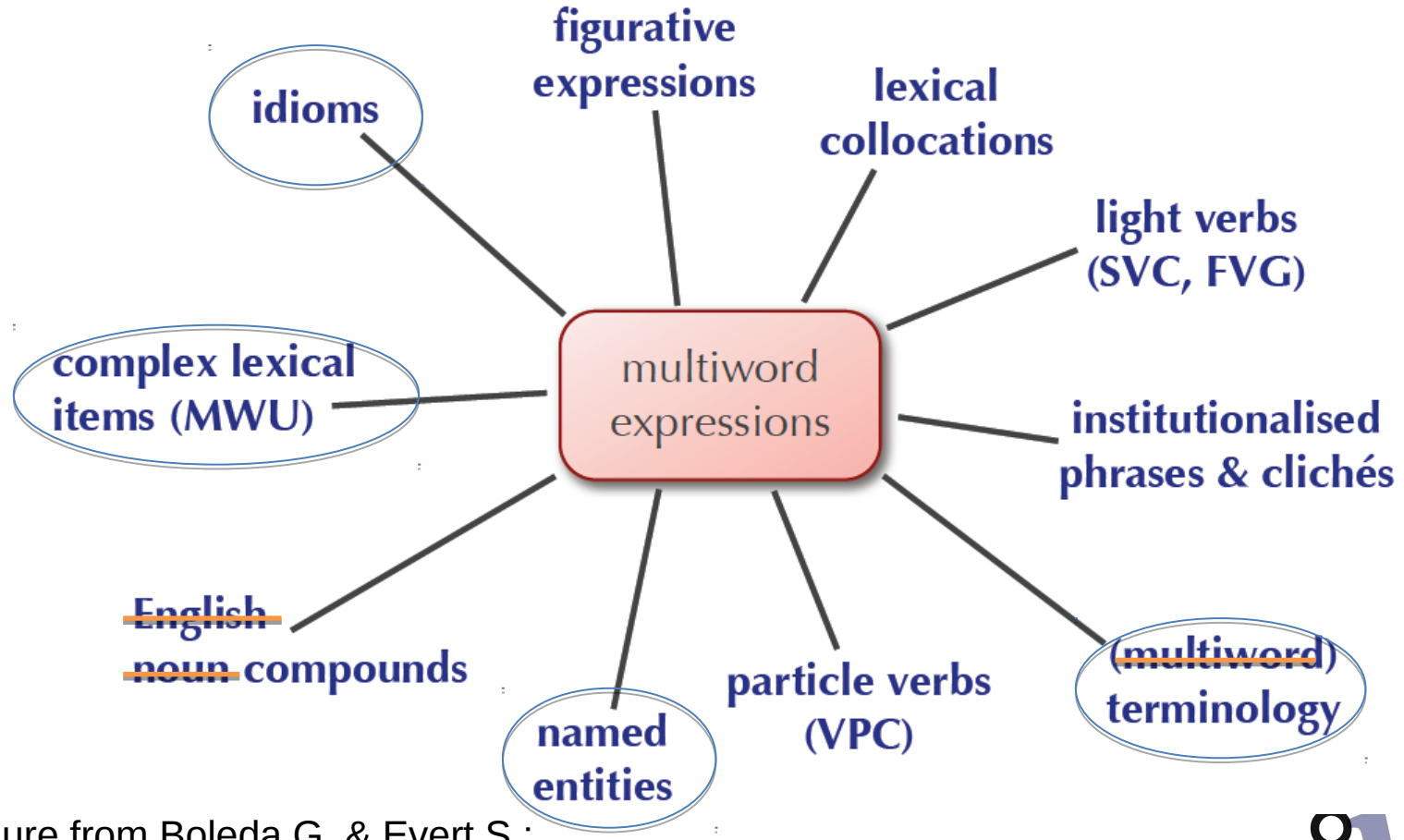
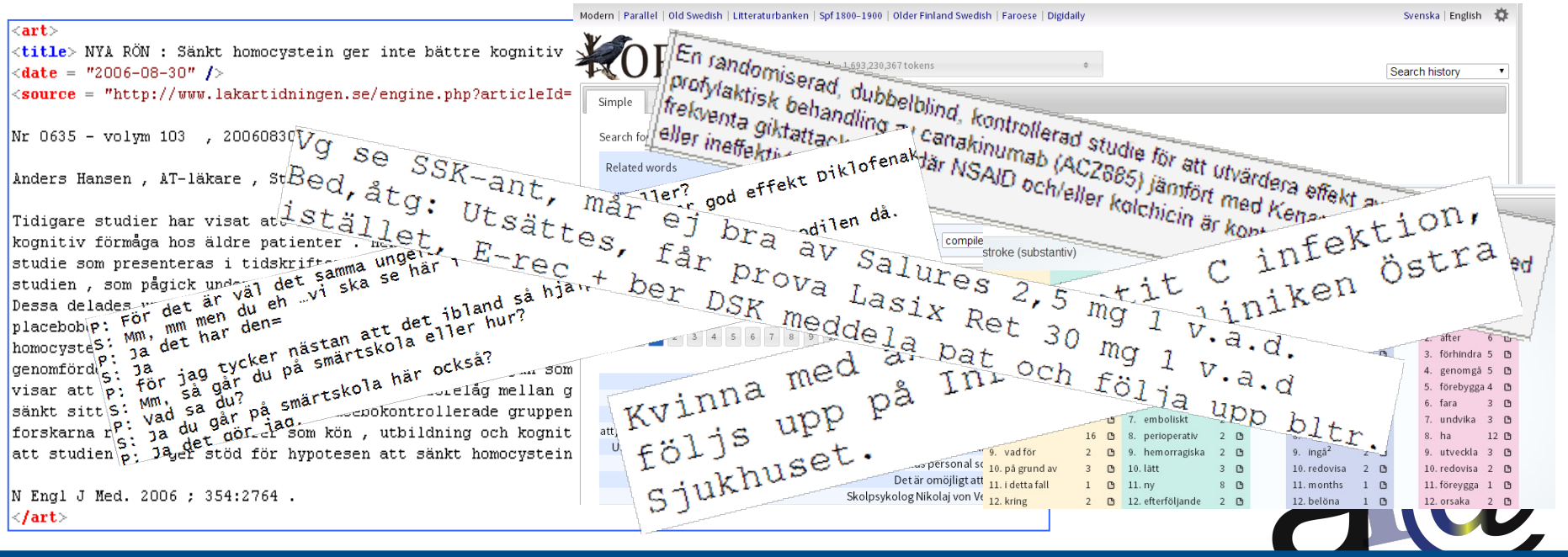


Figure from Boleda G. & Evert S.:
"Multiword Expressions: A pain in the neck of lexical semantics"

Textual Data – gets BIGger and BIGger

- Swedish medical corpora (various types)
  - scientific, clinical, guidelines, transcriptions …
- Enhance parsing by layers of pre-annotation
- Seemingly ”flat” annotation, but semantically rich
- Goal: Text mining, Relation/Event extraction, Semantic search



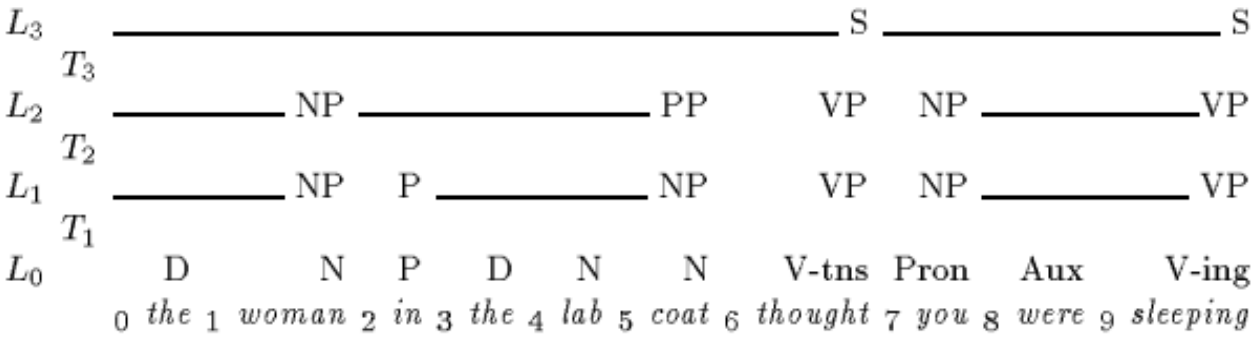
Related Work

- Sag I.A. et al. 2002. *Multiword expressions: A pain in the neck for NLP.*
- Leroy G. et al. 2003. *A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text.*
- Nivre J. and Nilsson J. 2004. *Multiword Units in Syntactic Parsing.*
- Lease M. and Charniak E. 2005. *Parsing Biomedical Literature*.
- Chou et al. 2006. *A Semi-Automatic Method for Annotating a Biomedical Proposition Bank*.
- Arun A. and Keller F. 2005. *Lexicalization in crosslinguistic probabilistic parsing: The case of French.*
- Buyko E. et al. 2009. *Event extraction from Trimmed Dependency Graphs*.
- Korkontzelos I. and Manandhar S. 2010. *Can Recognising Multiword Expressions Improve Shallow Parsing?*
- Constant M. et al. 2012. *Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing.*
- Constant M. et al. 2013. *Accounting for Contiguous Multiword Expressions in Shallow Parsing*
- …

Parsing Method in a Nutshell

Shallow Parsing – Finite-state cascades

- Sequences of levels (*Ln*) – phrases at one level are built on phrases at the previous level
- Parsing consists of a series of transductions (*Tn*) – spans of input elements are reduced to single elements in each transduction



- Each transduction is defined by a set of patterns (category + reg. expression) – each reg. expression is translated into a fs automaton, and union of all yields a single, deterministic fs level recognizer

Integrating Knowledge on MWE 1/3

- Function words – general vocabulary (612):
  - Prepositions, adverbs, conjunctions, determiners, pronouns
    - i_och_för_sig
    - på_grund_av
    - på_gott_och_ont
    - i_mångt_och_mycket
    - på_sätt_och_vis
    - på_det_hela_taget
    - i_grund_och_botten
    - till_och_med
    - …
- Lexicalised idioms (<IDIOM></IDIOM>)
  - <IDIOM>gå_ner_i_vikt</IDIOM>
  - <IDIOM>gå_in_i_väggen</IDIOM>
  - <IDIOM>gick_upp_i_vikt</IDIOM>
  - …

Integrating Knowledge on MWE 2/3

- Named entities (<ENAMEX>, <TIMEX>, <NUMEX>)
  - 8 main types of NEs; *person, location, organization, object/artifact, event, work, time* and *measure expressions*
    - <ENAMEX>Klassifikation av sjukdomar och hälsoproblem</ENAMEX>
    - <ENAMEX>Lagen om stöd och service</ENAMEX>
    - <ENAMEX>Svenska barnläkarföreningens arbetsgrupp för flyktingbarn</ENAMEX>
    - …
    - <TIMEX>i början av nästa år</TIMEX>
    - <TIMEX>så kort tid som möjligt</TIMEX>
    - <TIMEX>redan på ett tidigt stadium</TIMEX>
    - …
- Medical terminology
  - MeSH - Medical Subject Headings (<mesh></mesh>)
  - SNOMED-CT – Clinical Terms (<snomed></snomed>)
  - Drug names (<fass></fass>)

Integrating Knowledge on MWE 3/3

Length	# finding	Example
1	1886 (5.71%)	*flygfobi* flying phobia
2	**6407 (19.41%)**	*blå läppar* blue lips
3	5594 (16.95%)	*kramp i ben* leg kramp
4	5723 (17.34%)	*stel i flera leder* multiple stiff joints
5	4036 (12.22%)	*fenotyp Lu (b-)* Lu(b-) phenotype
6	3062 (9.27%)	*rädsla för att gå över gator* fear of crossing streets
7	2215 (6.71%)	*kan inte hålla kvar mat i munnen* unable to retain food in mouth
8	1291 (3.91%)	*distal resektionskant utan engagemang av carcinoma in situ* surgical distal margin uninvolved by in situ carcinoma
9	840 (2.54%)	*är inte beroende av hjälp för att kunna gå* independent walking
>9	1502 (4.55%)	*sen vaccin mot difteri, stelkramp, kikhosta , Haemophilus inluenza typ B och polio* did not attend 3rd DTP, Hib and polio vaccination

- SNOMED CT:
  - 33,001 findings
- avg length: 4,55 words
- only 5,71% of all *findings* consist of a single token

Integrating Knowledge on MWE

Sedan slutet av 1970-talet är det visat att den viktigaste orsaken till svårare diarré i samband med antibiotikabehandling är överväxt av bakterien Clostridium difficile .

Since the late 1970s, it is shown that the main cause of severe diarrhea associated with antibiotic therapy is the overgrowth of the bacterium Clostridium difficile.

MWE: Sedan slutet av 1970-talet är det visat att den viktigaste orsaken till svårare diarré <i_samband_med> antibiotikabehandling är överväxt av bakterien Clostridium difficile .

NER: <TIMEX TYPE="TME" SBT="DAT">Sedan slutet av 1970-talet</TIMEX> är det visat att den viktigaste orsaken till svårare diarré i_samband_med antibiotikabehandling är överväxt av bakterien <ENAMEX TYPE="TRM" SBT="MDO">Clostridium difficile</ENAMEX> .

TER: Sedan slutet av 1970-talet är det visat att den viktigaste orsaken till <snomed c="finding" h="409587002" o="svår diarré" f="inflection">svårare diarré</snomed> i_samband_med <snomed c="procedure" h="281789004" o="antibiotikabehandling" f="original">antibiotikabehandling</snomed> är överväxt av <snomed c="organism" h="409822003#41146007" o="bakterier" f="deletion-inflection-1">bakterien</snomed> <snomed c="organism" h="5933001" o="Clostridium difficile" f="original">Clostridium difficile</snomed> .

Integrating Knowledge on MWE

…the results from the MWE-processes are merged into a single representation format and fed into the syntactic analysis module > *super-chunking*

```
<s id="id.2">
 <t id="id.2_1"/> Sedan                  SPS              sedan                  TME/DAT-B  O       O
 <t id="id.2_2"/> slutet                 NCNSN@DS-VAL     slut                   TME/DAT-I  O       O
 <t id="id.2_3"/> av                     SPS              av                     TME/DAT-I  O       O
 <t id="id.2_4"/> 1970-talet             NCNSN@DS         1970-tal               TME/DAT-I  O       O
 <t id="id.2_5"/> är                     V@IPAS           vara                   O          O       O
 <t id="id.2_6"/> det                    PF@NS0@S         det                    O          O       O
 <t id="id.2_7"/> visat                  AF0NSNIS         visa                   O          O       O
 <t id="id.2_8"/> att                    CSS              att                    O          O       O
 <t id="id.2_9"/> den                    DF@US@S          den                    O          O       O
 <t id="id.2_10"/> viktigaste            AQS00NDS         viktig                 O          O       O
 <t id="id.2_11"/> orsaken               NCUSN@DS-VAL     orsak                  O          O       O
 <t id="id.2_12"/> till                  SPS              till                   O          O       O
 <t id="id.2_13"/> svårare               AQC00N0S         svår                   O          FND-B   O
 <t id="id.2_14"/> diarré                NCUSN@IS-VAL     diarré                 O          FND-I   C23-B
 <t id="id.2_15"/> i_samband_med         SPS              i_samband_med          O          O       O
 <t id="id.2_16"/> antibiotikabehandling NCUSN@IS         antibiotikabehandling  O          PRC-B   O
 <t id="id.2_17"/> är                    V@IPAS           vara                   O          O       O
 <t id="id.2_18"/> överväxt              NCUSN@IS-VAL     överväxt               O          O       G07-B
 <t id="id.2_19"/> av                    SPS              av                     O          O       O
 <t id="id.2_20"/> bakterien             NCUSN@DS         bakterie               O          ORG-B   B03-B
 <t id="id.2_21"/> Clostridium           XF               clostridium            MDO-B      ORG-B   B03-B
 <t id="id.2_22"/> difficile             XF               difficile              MDO-I      ORG-I   B03-I
 <t id="id.2_23"/> .                     FE               .                      O          O       O
</s>
```

Integrating Knowledge on MWE

På grund av typ 2 diabetes ökade utbredning i Kina , anordnade World Diabetes Fund tillsammans med Kinas hälsoministerium 2003-2008 en kampanj i landet .

Because of type 2 diabetes increased prevalence in China, organized the World Diabetes Fund, along with China's Ministry of Health from 2003 to 2008 a campaign in the country.

MWE: <På_grund_av> typ 2 diabetes ökade ...

NER: På_grund_av typ 2 diabetes ökade utbredning i <ENAMEX TYPE="LOC" SBT="PPL">Kina</ENAMEX> , anordnade <ENAMEX TYPE="ORG" SBT="FIN">World Diabetes Fund</ENAMEX> tillsammans med <ENAMEX TYPE="ORG" SBT="PLT">Kinas hälsoministerium</ENAMEX> <TIMEX TYPE="TME" SBT="DAT">2003-2008</TIMEX> en kampanj i landet .

TER: På_grund_av <snomed c="disorder" h="44054006" o="diabetes mellitus typ 2" f="modified">typ 2 diabetes</snomed> ökade utbredning i Kina , anordnade World <snomed c="disorder" h="73211009" o="diabetes" f="deletion">Diabetes</snomed1> Fund tillsammans med Kinas hälsoministerium 2003-2008 en kampanj i landet .

Integrating Knowledge on MWE

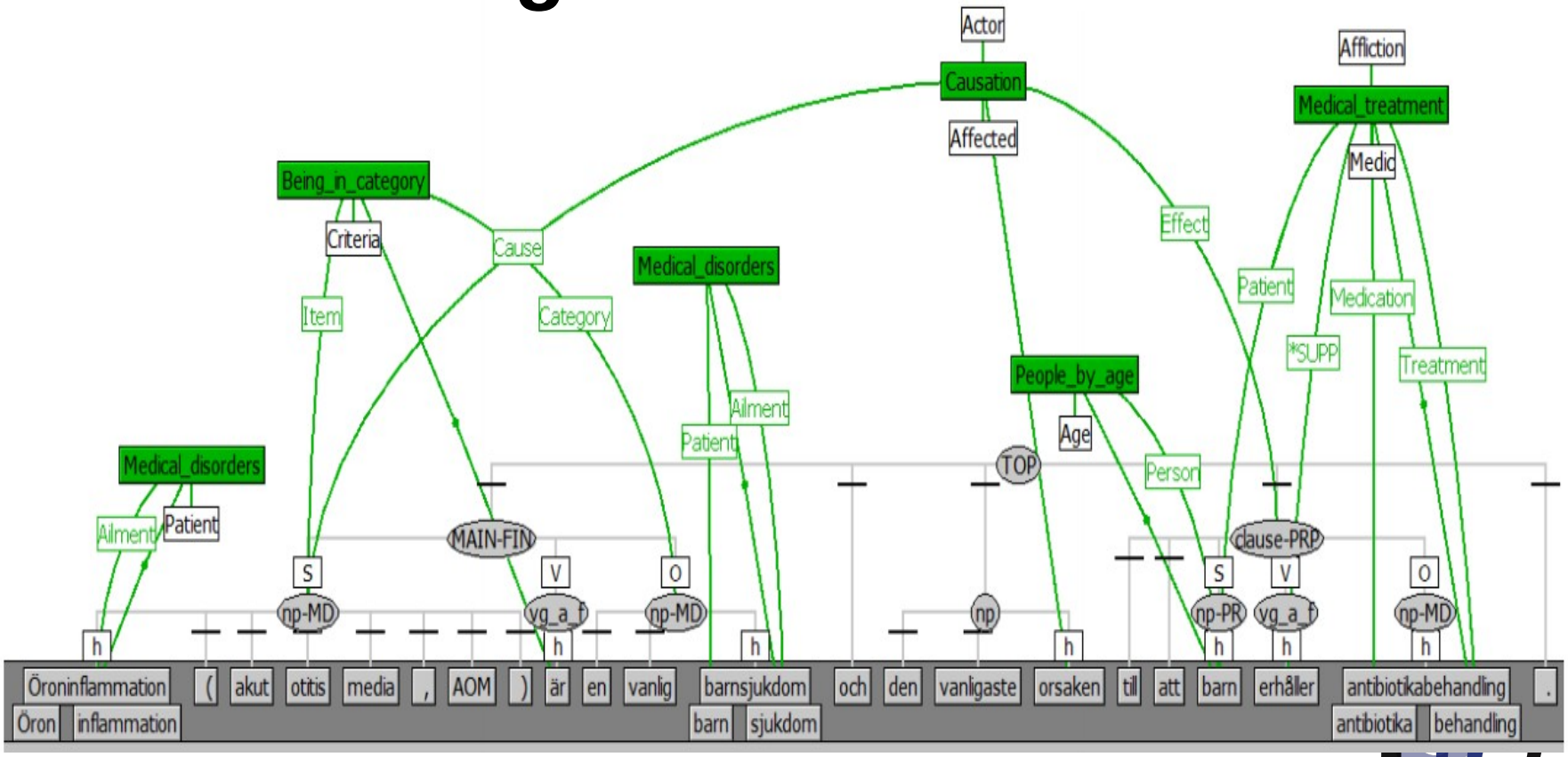
…the results from the MWE-processes are merged into a single representation format and fed into the syntactic analysis module > *super-chunking*

```
<s id="id.1">
 <t id="id.1_1"/> På_grund_av    SPS          på_grund_av     O          O       O
 <t id="id.1_2"/> typ            NCUSN@IS-NUMtyp              O          DIS-B   C18-B
 <t id="id.1_3"/> 2              MC00N0S      2               O          DIS-I   C18-I
 <t id="id.1_4"/> diabetes       NCUSN@IS     diabetes        O          DIS-I   C18-I
 <t id="id.1_5"/> ökade          AF00PN0S     öka             O          O       O
 <t id="id.1_6"/> utbredning     NCUSN@IS     utbredning      O          O       O
 <t id="id.1_7"/> i              SPS          i               O          O       O
 <t id="id.1_8"/> Kina           NP00N@0S     kina            LOC/PPL-B  O       Z01-B
 <t id="id.1_9"/> ,              FI           ,               O          O       O
 <t id="id.1_10"/> anordnade     V@IIAS       anordna         O          O       O
 <t id="id.1_11"/> World         NP00N@0S     world           ORG/FIN-B  O       O
 <t id="id.1_12"/> Diabetes      NCUSN@IS     diabetes        ORG/FIN-I  DIS-B   C18-B
 <t id="id.1_13"/> Fund          NP00N@0S     fund            ORG/FIN-I  O       N03-B
 <t id="id.1_14"/> tillsammans   RG0S-VAL     tillsammans     O          O       O
 <t id="id.1_15"/> med           SPS          med             O          O       O
 <t id="id.1_16"/> Kinas         NP00G@0S     kina            ORG/PLT-B  O       Z01-B
 <t id="id.1_17"/> hälsoministerium NCNSN@ISHälsoministeriumORG/PLT-I  O       O
 <t id="id.1_18"/> 2003-2008     MC00N0S      2003-2008       TME/DAT-B  O       O
 <t id="id.1_19"/> en            DI@US@S      en              O          O       O
 <t id="id.1_20"/> kampanj       NCUSN@IS     kampanj         O          O       O
 <t id="id.1_21"/> i             SPS          i               O          O       O
 <t id="id.1_22"/> landet        NCNSN@DS     land            O          O       O
 <t id="id.1_23"/> .             FE           .               O          O       O
</s>
```

Integrating Knowledge on MWE

```
...
  [MAIN-FIN
  V=[vg_a_f
    h=[V@IIAS  <t id="id.1_10"/> anordnade###anordna]]
  S=[orgn-entity
      [NP00N@0S-o  <t id="id.1_11"/> World###world]
      [NCUSN@IS-o  <t id="id.1_12"/> Diabetes###diabetes]
    h=[NP00N@0S-o  <t id="id.1_13"/> Fund###fund]]
    [rp_attach_pp
      [adv-general-val
        [RG0S-VAL  <t id="id.1_14"/> tillsammans###tillsammans]]
      [pp
        [SPS  <t id="id.1_15"/> med###med]
        [orgn-entity
          [NP00G@0S-o  <t id="id.1_16"/> Kinas###kina]
        h=[NCNSN@IS-o  <t id="id.1_17"/> hälsoministerium###hälsoministerium]]]]
    [time-entity
      [MC00N0S-Bt  <t id="id.1_18"/> 2003-2008###2003-2008]]
  O=[np3
      [DI@US@S  <t id="id.1_19"/> en###en]
    h=[NCUSN@IS  <t id="id.1_20"/> kampanj###kampanj]]
    [pp
...
```

Goal: SRL and Event Extraction using the Swedish FrameNet



SALTO: annotation of semantic roles

Results – Evaluation on a sample …

… based on the HO of phrases that signal grammatical subject and object. Despite errors at all levels of processing, some had limited impact for the grammatical relation extraction (e.g. pos errors), MWE had a positive effect in parsing – phrase recognition depends more on the shallow semantic annotation (precedence) than e.g. pos tagging

```
<s_id="id.223">
...
 <t id="id.223_11"/> några          PI@0P0@S        några       N/A-O
 <t id="id.223_12"/> av             SPS             av          N/A-O
 <t id="id.223_13"/> världens       NCUSG@DS        värld       N/A-O
 <t id="id.223_14"/> ledande        AP000N0S        ledande     N/A-O
 <t id="id.223_15"/> tidskrifter    NCUPN@IS        tidskrift   N/A-O
 <t id="id.223_16"/> (              FP              (           N/A-O
 <t id="id.223_17"/> N              NC000@0A        n           WRK/WAA-B
 <t id="id.223_18"/> Engl           NP00N@0S        engl        WRK/WAA-I
 <t id="id.223_19"/> J              NP00N@0S        j           WRK/WAA-I
 <t id="id.223_20"/> Med            SPS             med         WRK/WAA-I
 <t id="id.223_21"/> ,              FI              ,           N/A-O
 <t id="id.223_22"/> BMJ            NP00N@0S        bmj         WRK/WMD-B
 <t id="id.223_23"/> ,              FI              ,           N/A-O
 <t id="id.223_24"/> Lancet         NP00N@0S        lancet      WRK/WMD-B
 <t id="id.223_25"/> ,              FI              ,           N/A-O
...
</s>
```

	found	correct extracted	P	total available	R
subject	#1334	#1259	94.3%	#1298	96.9%
object	#638	#564	88.4%	#608	92.7%
indirect obj.	#13*	#2	15%	#4	50.0%

Precision: *# correct extracted relations/# total extracted relations*
Recall: *# correct extracted relations/# total available*

Future work – more forms of MWE

koord_redupl_adv

type	samordn
category	AdvP
definition	Upprepning av adverbial fungerar aspektuellt, eller påverkar aktionens innebörd i fråga om iteration, frekvens eller riktning.
structure	[Adv$_{=}$ och1 Adv$_{=}$ (och1 Adv$_{=}$)]
construction evoking elements	och[1]
common words	om[4] runt[2] igen[1]
internal construction elements	▪ Circumstance: cat=AdvP role=Circum ▪ Activity: cat=VP role=Activity ▪ Konj: lu=och[1] role=Konj
external construction elements	▪ Agent: cat=NP ▪ Patient: cat=NP ▪ Theme: cat=NP
examples	▪ Efter det [[körde]Activity [vi]Agent [r]koord_redupl_adv och testade varanc ▪ bjuder på [en härlig låt]Patient som [[spelas]Activity [om]Circumstance [och]cee [om]Circumstance]koord_redupl_adv igen här hemma . ▪ Sedan [[dök [det]Theme upp]Activity [igen]Circumstance [och]cee [igen]Circumstance]koord_redupl_adv och till slut tog jag upp det på ett APT möte . ▪ Och vem gillar inte [[en burk ärtsoppa]Activity [då]Circumstance [och]cee [då]Circumstance]koord_redupl_adv ?
reference	Lindström, Jan (1999): Vackert, vackert! Syntaktisk reduplikation i svenskan. Studier i nordisk filologi 77, Helsingfors.

investigate the addition of “new” types of suitable resources e.g. a *Swedish constructicon*: semi-productive, highly problematic for language technology; quite common; often neglected in both grammars and dictionaries e.g.
tautologies: “boys are boys”
repetitions: “very, very nice …”
…

<http://spraakbanken.gu.se/eng/swecxn>