



UNIVERSITY OF GOTHENBURG
DEPT OF SWEDISH

CLT

Språk
BANKEN



WG2: Improving Parsing Medical Discourse using Cascades of Multiword Expression Recognition

DIMITRIOS KOKKINAKIS

Dep. of Swedish, the Swedish Language Bank &
the Centre for Language Technology (CLT)

University of Gothenburg

Sweden

dimitrios.kokkinakis@svenska.gu.se



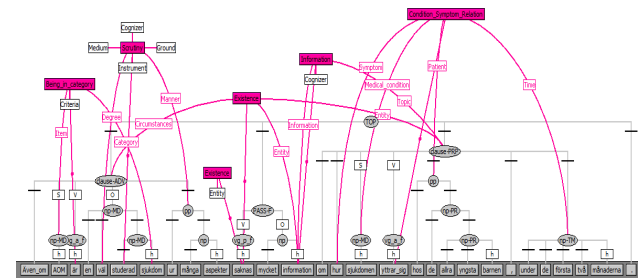


OUTLINE

...how a cascaded finite state constituent parser for Swedish behaves, with respect to parsing accuracy, where various types of MWE are automatically and successively recognized and introduced into the parser

```
<s id="id.149">
[MAIN-INF
[RGOA <t id="id.149_1"/> Bland_annot]
[vg_
h=[V@IPAS-AUX <t id="id.149_2"/> har]]
SBJ=[np-pronoun
h=[PF@NSO@s <t id="id.149_3"/> det]]
V=[vg_a_i
h=[V@IUAS <t id="id.149_4"/> visat_sig]]]
[CSS <t id="id.149_5"/> att]
[... ]
<s id="id.230">
[MAIN-FIN
SBJ=[np-entity-work
[NPOON@OS <t id="id.230_1"/> The]
[work-entity
[NPOON@OS-w <t id="id.230_2"/> Great]
[NPOON@OS-w <t id="id.230_3"/> Smoky]
[NPOON@OS-w <t id="id.230_4"/> Mountain]
h=[NPOON@OS-w <t id="id.230_5"/> Study]]]
V=[vg_a_f
h=[V@IPAS <t id="id.230_6"/> är]]
OBJ=[np
[DI@US@S <t id="id.230_7"/> en]
[AQPUSNIS <t id="id.230_8"/> amerikansk]
[AQPUSNIS <t id="id.230_9"/> longitudinell]
h=[NCUSN@IS <t id="id.230_10"/> befolkningsstudie]]]
[... ]
```

```
<s id="id.149">
<graph root="id.149_1">
<terminals>
<id id="id.149_3" word="Bland_annot" pos="NOUN" />
<id id="id.149_5" word="har" pos="AUX" />
<id id="id.149_7" word="det" pos="DET" />
<id id="id.149_9" word="visat" pos="AUX" />
<id id="id.149_11" word="att" pos="PART" />
<id id="id.149_14" word="konsumtion" pos="NOUN" />
<id id="id.149_16" word="av" pos="AUX" />
<id id="id.149_18" word="risa" pos="NOUN" />
<id id="id.149_20" word="är" pos="AUX" />
<id id="id.149_23" word="associerad" pos="AUX" />
<id id="id.149_25" word="med" pos="AUX" />
<id id="id.149_27" word="splitwords" pos="NOUN" />
<id id="id.149_29" word="Bland_annot" pos="NOUN" />
<id id="id.149_31" word="splitwords" pos="NOUN" />
<id id="id.149_33" word="splitwords" pos="NOUN" />
<id id="id.149_35" word="splitwords" pos="NOUN" />
<id id="id.149_37" word="splitwords" pos="NOUN" />
<id id="id.149_39" word="splitwords" pos="NOUN" />
<id id="id.149_41" word="splitwords" pos="NOUN" />
<id id="id.149_43" word="splitwords" pos="NOUN" />
<id id="id.149_45" word="splitwords" pos="NOUN" />
</terminals>
<frames>
<frame name="Obviousness" id="id.149_f1">
<target>
</target>
</frame>
<frame name="Ingestion" id="id.149_f2">
<target>
</target>
<fe name="Ingestibles" id="id.149_f2_e1">
<node idref="id.149_17"/>
</fe>
</frames>
<splitwords>
<splitword idref="id.149_3">
<part word="Bland" id="id.149_3_e0"/>
<part word="annot" id="id.149_3_e1"/>
</splitword>
<splitword idref="id.149_5">
<part word="visat" id="id.149_5_e0"/>
<part word="ar" id="id.149_5_e1"/>
</splitword>
<splitword idref="id.149_7">
<part word="visat" id="id.149_7_e0"/>
<part word="ar" id="id.149_7_e1"/>
</splitword>
<splitword idref="id.149_9">
<part word="visat" id="id.149_9_e0"/>
<part word="ar" id="id.149_9_e1"/>
</splitword>
<splitword idref="id.149_11">
<part word="visat" id="id.149_11_e0"/>
<part word="ar" id="id.149_11_e1"/>
</splitword>
<splitword idref="id.149_14">
<part word="visat" id="id.149_14_e0"/>
<part word="ar" id="id.149_14_e1"/>
</splitword>
<splitword idref="id.149_16">
<part word="visat" id="id.149_16_e0"/>
<part word="ar" id="id.149_16_e1"/>
</splitword>
<splitword idref="id.149_18">
<part word="visat" id="id.149_18_e0"/>
<part word="ar" id="id.149_18_e1"/>
</splitword>
<splitword idref="id.149_20">
<part word="visat" id="id.149_20_e0"/>
<part word="ar" id="id.149_20_e1"/>
</splitword>
<splitword idref="id.149_23">
<part word="visat" id="id.149_23_e0"/>
<part word="ar" id="id.149_23_e1"/>
</splitword>
<splitword idref="id.149_25">
<part word="visat" id="id.149_25_e0"/>
<part word="ar" id="id.149_25_e1"/>
</splitword>
<splitword idref="id.149_27">
<part word="visat" id="id.149_27_e0"/>
<part word="ar" id="id.149_27_e1"/>
</splitword>
<splitword idref="id.149_29">
<part word="visat" id="id.149_29_e0"/>
<part word="ar" id="id.149_29_e1"/>
</splitword>
<splitword idref="id.149_31">
<part word="visat" id="id.149_31_e0"/>
<part word="ar" id="id.149_31_e1"/>
</splitword>
<splitword idref="id.149_33">
<part word="visat" id="id.149_33_e0"/>
<part word="ar" id="id.149_33_e1"/>
</splitword>
<splitword idref="id.149_35">
<part word="visat" id="id.149_35_e0"/>
<part word="ar" id="id.149_35_e1"/>
</splitword>
<splitword idref="id.149_37">
<part word="visat" id="id.149_37_e0"/>
<part word="ar" id="id.149_37_e1"/>
</splitword>
<splitword idref="id.149_39">
<part word="visat" id="id.149_39_e0"/>
<part word="ar" id="id.149_39_e1"/>
</splitword>
<splitword idref="id.149_41">
<part word="visat" id="id.149_41_e0"/>
<part word="ar" id="id.149_41_e1"/>
</splitword>
<splitword idref="id.149_43">
<part word="visat" id="id.149_43_e0"/>
<part word="ar" id="id.149_43_e1"/>
</splitword>
<splitword idref="id.149_45">
<part word="visat" id="id.149_45_e0"/>
<part word="ar" id="id.149_45_e1"/>
</splitword>
</splitwords>
</graph>
grund_av 0 0 DIS-B 0
p 0 0 DIS-I 0
diabetes 0 0 DIS-I 0
ja 0 0 0 0
<id id="id.149_47"> grund_av 0 0 0 0
<id id="id.149_49"> p 0 0 0 0
<id id="id.149_51"> diabetes 0 0 0 0
<id id="id.149_53"> ja 0 0 0 0
<id id="id.149_55"> grund_av 0 0 0 0
<id id="id.149_57"> kina 0 0 0 0
<id id="id.149_59"> kina 0 0 0 0
<id id="id.149_61"> anordnade 0 0 0 0
<id id="id.149_63"> World 0 0 0 0
<id id="id.149_65"> Diabetes 0 0 0 0
<id id="id.149_67"> Fund 0 0 0 0
<id id="id.149_69"> tillsammans 0 0 0 0
<id id="id.149_71"> med 0 0 0 0
<id id="id.149_73"> Kinas 0 0 0 0
<id id="id.149_75"> halsoministerium 0 0 0 0
<id id="id.149_77"> 2003-2008 0 0 0 0
<id id="id.149_79"> en 0 0 0 0
<id id="id.149_81"> kampanj 0 0 0 0
<id id="id.149_83"> SPS 0 0 0 0
<id id="id.149_85"> landet 0 0 0 0
<id id="id.149_87"> FE 0 0 0 0
</>
```



	Precision	Recall	F-score
i*	77.92%	76.79%	77.35%
iii*	77.96%	76.82%	77.39%
iv*	80.71%	79.93%	80.32%
v*	86.67%	87.10%	86.89%

