The Legato annotation tool

David Alfter, Therese Lindström Tiedemann, Elena Volodina

SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

Acknowledgements

We thank the L2 profiles project

The project is financed by *Riksbankens Jubileumsfond* during years 2018-2020 through a grant P17-0716:1

This project intends to study the development of lexical and grammatical competences in L2 learners of Swedish.

https://spraakbanken.gu.se/eng/l2-profiles

SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

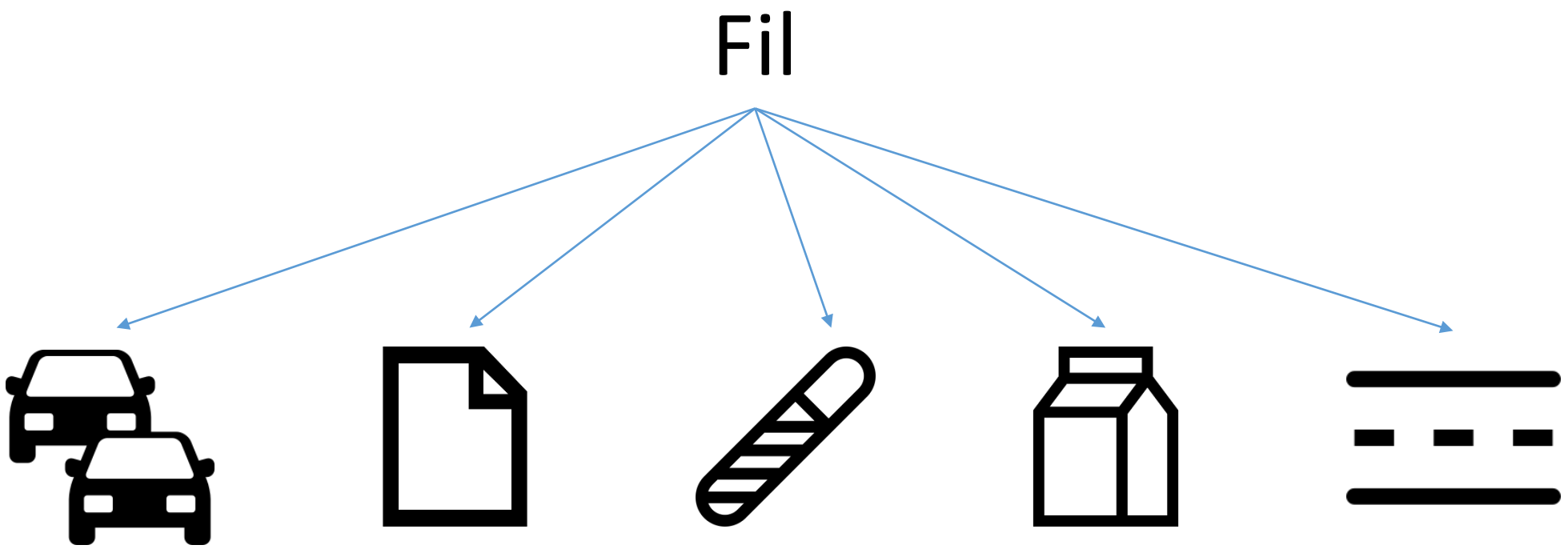
From lemmas to senses

- SVALex (François et al., 2016)
- SweLLex (Volodina et al., 2016)

→ SenSVALex

→ SenSweLLex

Why we need sense-based word lists



Idea

- SB has a plethora of resources

→ Use existing resources to enrich sense-based word lists

- Augment with manual annotation

→ GUI for annotators

Data preparation

- Calculation of SenSVALex and SenSweLLex

Creating sense-based word lists

- Analogous to CEFRLex resources (http://cental.uclouvain.be/cefrlex/)

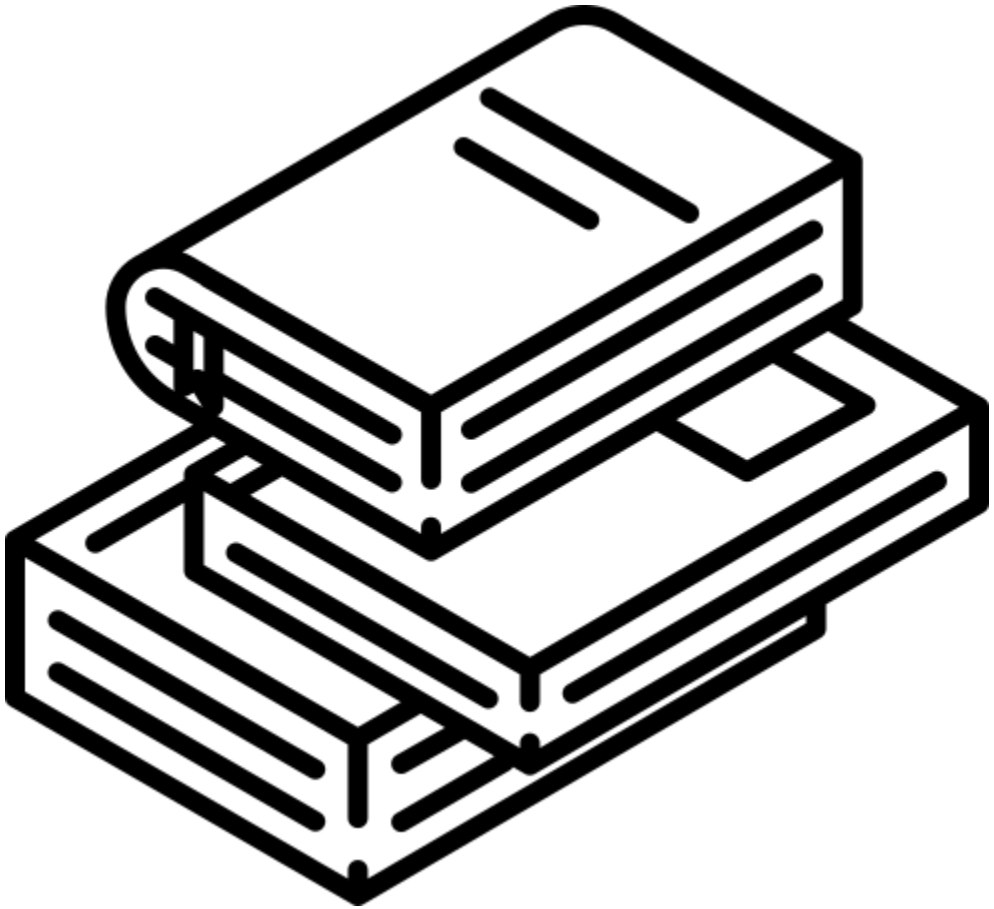
Creating sense-based word lists

COCTAILL corpus (Volodina et al., 2014)

SweLL corpus (Volodina et al., 2016)

- Automatic annotation through Sparv (Borin et al., 2016)
- Extract lemgram+sense+pos tuples with frequencies per level
- Map to target level

The corpus



SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

Automatic annotation



SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

```
<sentence id="d0fc0ca-d07cf63" _geocontext="|">
<w pos="VB" msd="VB.PRS.AKT" lemma="|kunna|" lex="|kunna..vb.1|" sense="|kunna..2:0.449|kunna..3:0.225|kunna..1:0.200|kunna..4:0.127|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="01" deprel="ROOT" blingbring="|förmå
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|du|" lex="|du..pn.1|" sense="|du..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="02" dephead="01" deprel="SS" blingbring="|" swefn="|" sentiment="0.0578" sentimen
<w pos="VB" msd="VB.INF.AKT" lemma="|gapa|" lex="|gapa..vb.1|" sense="|gapa..1:0.739|gapa..2:0.162|gapa..3:0.057|gapa..4:0.042|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="03" dephead="01" deprel="VG" blingbring="|
<w pos="AB" msd="AB" lemma="|stor|" lex="|stor..av.1|" sense="|stor..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="04" dephead="05" deprel="AA" blingbring="|betydenhet|grad|jämförelse|medelmåttighet|musik|m
<w pos="AB" msd="AB" lemma="|så|" lex="|så..ab.1|" sense="|så..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="05" dephead="03" deprel="AA" blingbring="|metod|omständighet|" swefn="|Sufficiency|" sentiment="0
<w pos="VB" msd="VB.PRS.AKT" lemma="|skola|" lex="|skola..vb.2|" sense="|skola..4:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="06" dephead="03" deprel="VG" blingbring="|beredelse|bestämmelse|bundenhet|förfar
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|jag|" lex="|jag..pn.1|" sense="|jag..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="07" dephead="08" deprel="SS" blingbring="|" swefn="|" sentiment="0.6341" senti
<w pos="VB" msd="VB.INF.AKT" lemma="|titta|" lex="|titta..vb.1|" sense="|titta..1:-1.000|" prefix="|titt..nn.1|tita..nn.1|titta..vb.1|" suffix="|ta..vb.1|" compwf="|titt+ta|tit+ta|" complemgram="|titt..nn.1+ta..vb.1:1.166e-10
<w pos="PP" msd="PP" lemma="|i|" lex="|i..pp.1|" sense="|i..2:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="09" dephead="08" deprel="RA" blingbring="|" swefn="|" sentiment="0.836" sentimentclass="neutral">i</
<w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|hals|" lex="|hals..nn.1|" sense="|hals..1:0.953|hals..2:0.047|" prefix="|hala..vb.1|hal..av.1|" suffix="|s..nn.1|" compwf="|hal+sen|" complemgram="|hal..av.1+s..nn.1:1.010e-20|hala
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="11" dephead="01" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
<sentence id="d03eda4-d06a9c4" _geocontext="|">
<w pos="VB" msd="VB.IMP.AKT" lemma="|säga|" lex="|säga..vb.1|" sense="|säga..1:0.588|säga..3:0.401|säga..2:0.011|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" deprel="ROOT" blingbring="|meddelande|tal|" swefn="|S
<w pos="IN" msd="IN" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="1" deprel="OO" blingbring="|" swefn="|">aaaaa</w>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" dephead="1" deprel="IG" blingbring="|" swefn="|">...</w>
</sentence>
<sentence id="d0f8142-d048caa" _geocontext="|">
<ne ex="ENAMEX" type="PRS" subtype="HUM" name="Daniel">
<w pos="PM" msd="PM.NOM" lemma="|Daniel|" lex="|Daniel..pm.1|" sense="|Daniel..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" deprel="ROOT" blingbring="|" swefn="|" sentiment="0.4104" sentimentclass="neut
</ne>
<w pos="MID" msd="MID" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="1" deprel="IQ" blingbring="|" swefn="|">:</w>
<w pos="IN" msd="IN" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" dephead="1" deprel="ROOT" blingbring="|" swefn="|">Aaaaa</w>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4" dephead="3" deprel="IG" blingbring="|" swefn="|">...</w>
</sentence>
<sentence id="d06c01c-d06dade" _geocontext="|">
<w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|läkare|" lex="|läkare..nn.1|" sense="|läkare..1:-1.000|" prefix="|lä..nn.1|läka..vb.1|" suffix="|kar..nn.1|ar..nn.1|" compwf="|lä+karen|läk+aren|" complemgram="|lä..nn.1+kar..nn.1:
<w pos="MID" msd="MID" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="1" deprel="IQ" blingbring="|" swefn="|">:</w>
<ne ex="ENAMEX" type="PRS" subtype="HUM" name="Bra">
<w pos="JJ" msd="JJ.POS.UTR+NEU.SIN+PLU.IND+DEF.NOM" lemma="|bra|" lex="|bra..av.2|" sense="|bra..4:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" deprel="ROOT" blingbring="|fördel|gott|jämförelse|" swefn="
</ne>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4" dephead="3" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
<sentence id="d0c9a2f-d0cc408" _geocontext="|">
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|jag|" lex="|jag..pn.1|" sense="|jag..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" dephead="2" deprel="SS" blingbring="|" swefn="|" sentiment="0.6341" sentime
<w pos="VB" msd="VB.PRS.AKT" lemma="|vilja|" lex="|vilja..vb.1|" sense="|vilja..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" deprel="ROOT" blingbring="|syfte|vilja|villighet|" swefn="|Desiring|" sentime
<w pos="VB" msd="VB.INF.AKT" lemma="|lyssna|" lex="|lyssna..vb.1|" sense="|lyssna..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" dephead="2" deprel="VG" blingbring="|hörsel|kunskapsbegär|samtycke|uppmärk
<w pos="PP" msd="PP" lemma="|på|" lex="|på..pp.1|" sense="|på..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4" dephead="3" deprel="OA" blingbring="|" swefn="|" sentiment="0.6675" sentimentclass="neutral">p
<w pos="NN" msd="NN.UTR.PLU.DEF.NOM" lemma="|lunga|" lex="|lunga..nn.1|" sense="|lunga..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="5" dephead="4" deprel="PA" blingbring="|luftrör|luftström|talröst|uppvak
<w pos="AB" msd="AB" lemma="|också|" lex="|också..ab.1|" sense="|också..1:-1.000|" prefix="|ock..ab.1|" suffix="|så..ab.1|" compwf="|" complemgram="|" ref="6" dephead="3" deprel="+A" blingbring="|vidfogning|" swefn="|" sentim
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="7" dephead="2" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
<sentence id="d02a0ec-d02b30e" _geocontext="|">
<w pos="VB" msd="VB.PRS.AKT" lemma="|kunna|" lex="|kunna..vb.1|" sense="|kunna..2:0.461|kunna..3:0.333|kunna..1:0.116|kunna..4:0.090|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" deprel="ROOT" blingbring="|förmåg
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|du|" lex="|du..pn.1|" sense="|du..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="3" deprel="SS" blingbring="|" swefn="|" sentiment="0.0578" sentimentc
<w pos="VB" msd="VB.INF.AKT" lemma="|ta|ta av|ta sig|" lex="|ta..vb.1|ta_av..vbm.1|ta_sig..vbm.1|" sense="|ta..3:0.479|ta..2:0.296|ta..1:0.193|ta..4:0.031|ta_av..1:-1.000|ta_av..2:-1.000|ta_av..3:-1.000|ta_av..4:-1.000|ta_sig
<w pos="PL" msd="PL" lemma="|av|ta av:3|" lex="|av..ab.1|ta_av..vbm.1:3|" sense="|av..3:-1.000|ta_av..1:3:-1.000|ta_av..2:3:-1.000|ta_av..3:3:-1.000|ta_av..4:3:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4"
<w pos="PN" msd="PN.UTR.SIN.DEF.OBJ" lemma="|du|ta sig:3|" lex="|du..pn.1|ta_sig..vbm.1:3|" sense="|du..1:-1.000|ta_sig..1:3:-1.000|ta_sig..2:3:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="5" dephead="3" dep
<w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|skjorta|" lex="|skjorta..nn.1|" sense="|skjorta..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="6" dephead="3" deprel="OO" blingbring="|påklädsel|" swefn="|Clothi
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="7" dephead="1" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
```


SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

```
<sentence id="d0fc0ca-d07cf63" _geocontext="|">
<w pos="VB" msd="VB.PRS.AKT" lemma="|kunna|" lex="|kunna..vb.1|" sense="|kunna..2:0.449|kunna..3:0.225|kunna..1:0.200|kunna..4:0.127|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="01" deprel="ROOT" blingbring="|förmå
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|du|" lex="|du..pn.1|" sense="|du..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="02" dephead="01" deprel="SS" blingbring="|" swefn="|" sentiment="0.0578" sentimen
<w pos="VB" msd="VB.INF.AKT" lemma="|gapa|" lex="|gapa..vb.1|" sense="|gapa..1:0.739|gapa..2:0.162|gapa..3:0.057|gapa..4:0.042|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="03" dephead="01" deprel="VG" blingbring="|
<w pos="AB" msd="AB" lemma="|stor|" lex="|stor..av.1|" sense="|stor..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="04" dephead="05" deprel="AA" blingbring="|betydenhet|grad|jämförelse|medelmåttighet|musik|m
<w pos="AB" msd="AB" lemma="|så|" lex="|så..ab.1|" sense="|så..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="05" dephead="03" deprel="AA" blingbring="|metod|omständighet|" swefn="|Sufficiency|" sentiment="0
<w pos="VB" msd="VB.PRS.AKT" lemma="|skola|" lex="|skola..vb.2|" sense="|skola..4:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="06" dephead="03" deprel="VG" blingbring="|beredelse|bestämmelse|bundenhet|förfar
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|jag|" lex="|jag..pn.1|" sense="|jag..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="07" dephead="08" deprel="SS" blingbring="|" swefn="|" sentiment="0.6341" senti
<w pos="VB" msd="VB.INF.AKT" lemma="|titta|" lex="|titta..vb.1|" sense="|titta..1:-1.000|" prefix="|titt..nn.1|tita..nn.1|titta..vb.1|" suffix="|ta..vb.1|" compwf="|titt+ta|tit+ta|" complemgram="|titt..nn.1+ta..vb.1:1.166e-10
<w pos="PP" msd="PP" lemma="|i|" lex="|i..pp.1|" sense="|i..2:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="09" dephead="08" deprel="RA" blingbring="|" swefn="|" sentiment="0.836" sentimentclass="neutral">i</
<w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|hals|" lex="|hals..nn.1|" sense="|hals..1:0.953|hals..2:0.047|" prefix="|hala..vb.1|hal..av.1|" suffix="|s..nn.1|" compwf="|hal+sen|" complemgram="|hal..av.1+s..nn.1:1.010e-20|hala
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="11" dephead="01" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
<sentence id="d03eda4-d06a9c4" _geocontext="|">
<w pos="VB" msd="VB.IMP.AKT" lemma="|säga|" lex="|säga..vb.1|" sense="|säga..1:0.588|säga..3:0.401|säga..2:0.011|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" deprel="ROOT" blingbring="|meddelande|tal|" swefn="|S
<w pos="IN" msd="IN" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="1" deprel="OO" blingbring="|" swefn="|">aaaaa</w>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" dephead="1" deprel="IG" blingbring="|" swefn="|">...</w>
</sentence>
<sentence id="d0f8142-d048caa" _geocontext="|">
<ne ex="ENAMEX" type="PRS" subtype="HUM" name="Daniel">
<w pos="PM" msd="PM.NOM" lemma="|Daniel|" lex="|Daniel..pm.1|" sense="|Daniel..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" deprel="ROOT" blingbring="|" swefn="|" sentiment="0.4104" sentimentclass="neut
</ne>
<w pos="MID" msd="MID" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="1" deprel="IQ" blingbring="|" swefn="|">:</w>
<w pos="IN" msd="IN" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" dephead="1" deprel="ROOT" blingbring="|" swefn="|">Aaaaa</w>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4" dephead="3" deprel="IG" blingbring="|" swefn="|">...</w>
</sentence>
<sentence id="d06c01c-d06dade" _geocontext="|">
<w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|läkare|" lex="|läkare..nn.1|" sense="|läkare..1:-1.000|" prefix="|lä..nn.1|läka..vb.1|" suffix="|kar..nn.1|ar..nn.1|" compwf="|lä+karen|läk+aren|" complemgram="|lä..nn.1+kar..nn.1:
<w pos="MID" msd="MID" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="1" deprel="IQ" blingbring="|" swefn="|">:</w>
<ne ex="ENAMEX" type="PRS" subtype="HUM" name="Bra">
<w pos="JJ" msd="JJ.POS.UTR+NEU.SIN+PLU.IND+DEF.NOM" lemma="|bra|" lex="|bra..av.2|" sense="|bra..4:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" deprel="ROOT" blingbring="|fördel|gott|jämförelse|" swefn="
</ne>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4" dephead="3" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
<sentence id="d0c9a2f-d0cc408" _geocontext="|">
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|jag|" lex="|jag..pn.1|" sense="|jag..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" dephead="2" deprel="SS" blingbring="|" swefn="|" sentiment="0.6341" sentime
<w pos="VB" msd="VB.PRS.AKT" lemma="|vilja|" lex="|vilja..vb.1|" sense="|vilja..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" deprel="ROOT" blingbring="|syfte|vilja|villighet|" swefn="|Desiring|" sentime
<w pos="VB" msd="VB.INF.AKT" lemma="|lyssna|" lex="|lyssna..vb.1|" sense="|lyssna..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="3" dephead="2" deprel="VG" blingbring="|hörsel|kunskapsbegär|samtycke|uppmärk
<w pos="PP" msd="PP" lemma="|på|" lex="|på..pp.1|" sense="|på..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4" dephead="3" deprel="OA" blingbring="|" swefn="|" sentiment="0.6675" sentimentclass="neutral">p
<w pos="NN" msd="NN.UTR.PLU.DEF.NOM" lemma="|lunga|" lex="|lunga..nn.1|" sense="|lunga..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="5" dephead="4" deprel="PA" blingbring="|luftrör|luftström|talröst|uppvak
<w pos="AB" msd="AB" lemma="|också|" lex="|också..ab.1|" sense="|också..1:-1.000|" prefix="|ock..ab.1|" suffix="|så..ab.1|" compwf="|" complemgram="|" ref="6" dephead="3" deprel="+A" blingbring="|vidfogning|" swefn="|" sentim
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="7" dephead="2" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
<sentence id="d02a0ec-d02b30e" _geocontext="|">
<w pos="VB" msd="VB.PRS.AKT" lemma="|kunna|" lex="|kunna..vb.1|" sense="|kunna..2:0.461|kunna..3:0.333|kunna..1:0.116|kunna..4:0.090|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="1" deprel="ROOT" blingbring="|förmåg
<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|du|" lex="|du..pn.1|" sense="|du..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="2" dephead="3" deprel="SS" blingbring="|" swefn="|" sentiment="0.0578" sentimentc
<w pos="VB" msd="VB.INF.AKT" lemma="|ta|ta av|ta sig|" lex="|ta..vb.1|ta_av..vbm.1|ta_sig..vbm.1|" sense="|ta..3:0.479|ta..2:0.296|ta..1:0.193|ta..4:0.031|ta_av..1:-1.000|ta_av..2:-1.000|ta_av..3:-1.000|ta_av..4:-1.000|ta_sig
<w pos="PL" msd="PL" lemma="|av|ta av:3|" lex="|av..ab.1|ta_av..vbm.1:3|" sense="|av..3:-1.000|ta_av..1:3:-1.000|ta_av..2:3:-1.000|ta_av..3:3:-1.000|ta_av..4:3:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="4"
<w pos="PN" msd="PN.UTR.SIN.DEF.OBJ" lemma="|du|ta sig:3|" lex="|du..pn.1|ta_sig..vbm.1:3|" sense="|du..1:-1.000|ta_sig..1:3:-1.000|ta_sig..2:3:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="5" dephead="3" dep
<w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|skjorta|" lex="|skjorta..nn.1|" sense="|skjorta..1:-1.000|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="6" dephead="3" deprel="OO" blingbring="|påklädsel|" swefn="|Clothi
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|" prefix="|" suffix="|" compwf="|" complemgram="|" ref="7" dephead="1" deprel="IP" blingbring="|" swefn="|">.</w>
</sentence>
```


SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

A sense-based resource

Lemgram	Sense	POS	A1	A2	B1	B2	C1
Kunna..vb.1	Kunna..1	VB	12	1	7	13	0
Kunna..vb.1	Kunna..2	VB	0	0	3	2	4
Kunna..vb.1	Kunna..3	VB	0	25	23	12	9
Kunna..vb.1	Kunna..4	VB	32	12	0	0	0

A sense-based graded resource

Lemgram	Sense	POS	Target
Kunna..vb.1	Kunna..1	VB	A1
Kunna..vb.1	Kunna..2	VB	B2
Kunna..vb.1	Kunna..3	VB	A2
Kunna..vb.1	Kunna..4	VB	A1


But wait...
How good is WSD?

But wait…

- How good is WSD?

- Manual check on 15 textbook texts
  - Assistants in Finland (linguists) in L2P
  - 3 texts per level

But wait…

- How good is WSD?

- Baseline: always assign first sense
  - 77.36% accuracy
- Current WSD:
  - 88.74% accuracy

Data linking

- Saldo and Saldo's morphology
  - Nominal gender
  - Nominal declension
  - Verbal conjugation
  - Adjectival declension → derived from given forms in SaldoM
  - …
- Swesaurus
  - Synonyms

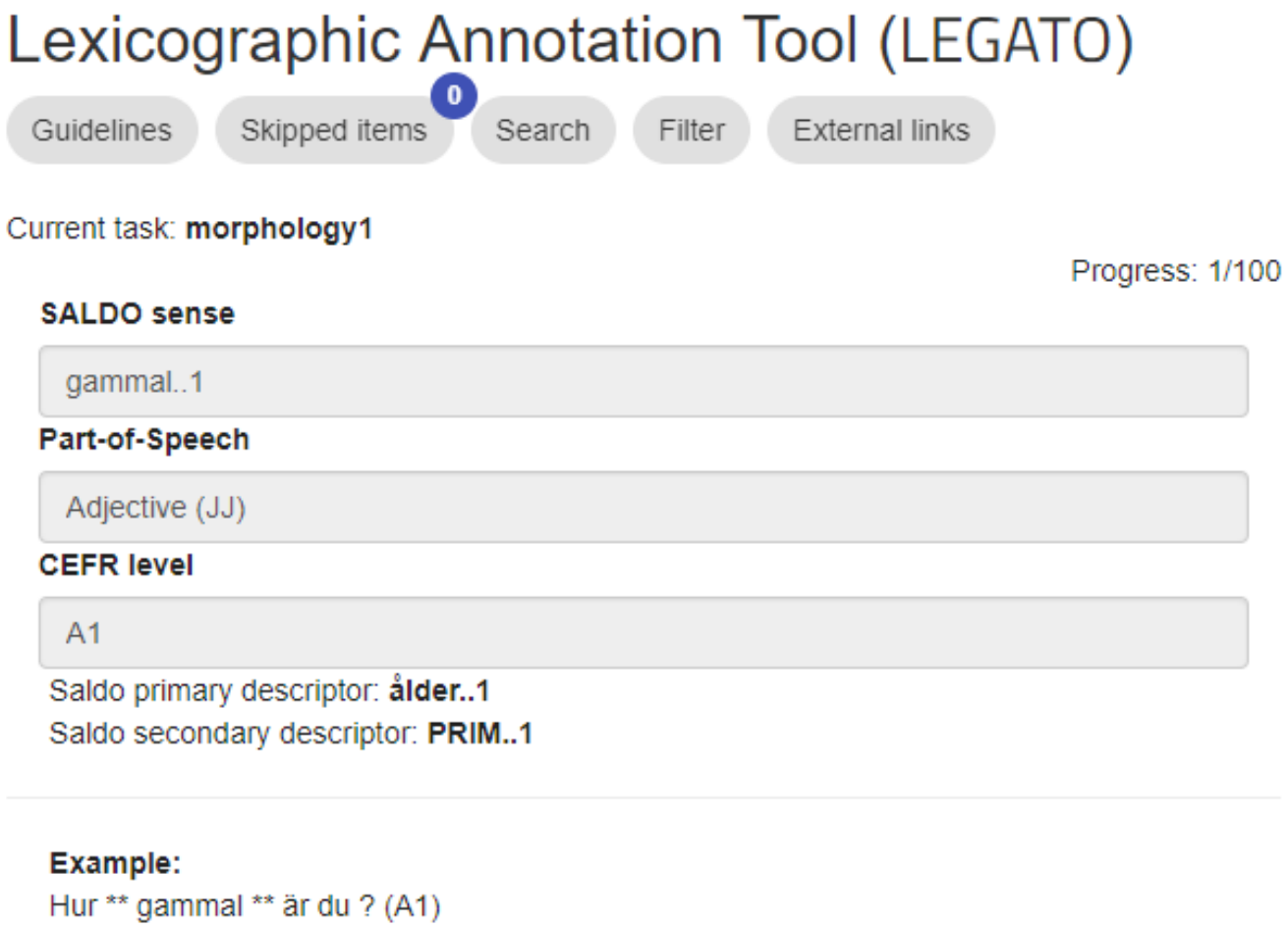
Data linking – Possible future resources

- SAOL
  - Transitivity
- Lexin/COCTAILL
  - Topical information

Categories for annotation

Aspect	Explanation / choices	Mode	Resources for auto-enrichment
1 Adj/adv structure	comparisons: periphr.: *(mer/mest) entusiastisk*; morph.: *vacker-vackrare-vackrast*; irreg.: *god-bra-bäst*	A-M[2]	Saldo-Morphology
2 Adj declension	decl. 1 & 2, irregular, indeclinable	A-M	Saldo-Morphology
3 Morphology 1	word analysis for morphemes: *oändlig*: prefix:*o-*; root:*-änd-*; suffix:*-lig*	M[3]	
4 Morphology 2	word-building: root, compound, derivation, suppletion, lexicalized, MWE[1]	M	
5 MWE type	taxonomy under development	M	
6 Nom declension	decl. 1-6, extra	A[4]	Saldo-Morphology
7 Nom gender	common, neuter, both, N/A	A	Saldo-Morphology
8 Nom type	abstract–concrete, (un)countable, (non)collective, (in)animate, proper name, unit of measurement	M	
9 Register	neutral, formal, informal, sensitive	M	
10 Synonyms	free input, same word class	A-M	Swesaurus
11 Topics/domains	general + 40 CEFR-related topics[5]	A-M	Lexin, COCTAILL
12 Transitivity	(in-, di-)transitive, N/A	A-M	SAOL (under negotiation)
13 Verb category	lexical, modal, auxiliary, copula, reciprocal, deponent	M	
14 Verb conjugation	conjugations 1-4, irregular, N/A	A	Saldo-Morphology
15 Verb action type	motion, state, punctual, process[6]	M	

The tool v.1



SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

The guidelines

- Extensive document (30 pages excluding appendix)

Piloting the tool

- 5 CEFR levels
- 4 word classes (nouns, verbs, adjectives, adverbs)
- 5 items per word class and CEFR level

➔ 5 * 4 * 5 = 100 items

Piloting results

- Inter-annotator agreement
  - Automatic analysis and Therese (IAA 1)
  - Therese and Elena (IAA 2)

Piloting results

Category	IAA 1	IAA 2
nominal declension (6)	0.85	0.80
nominal gender (7)	0.82	0.73
nominal type (5)		0.20
verbal conjugation (14)	0.82	0.94
adjectival declension (2)	0.49	
adjectival adverbial structure (1)	0.39	
morphology 1 (3)		0.48
Overall κ	0.73	0.60

Piloting results

Based on IAA results, certain categories were chosen to be not included in the tool since the automatic annotation was deemed good enough

- Nominal declension
- Nominal gender
- Verbal conjugation

Piloting results – Observations

- Closed-answer single-answer categories have higher IAA
- Open-answer and multiple choice categories have lower IAA

Exception:

- Adjectival declension
- Possibly due to automatic (regular) paradigm expansion

Automatic paradigm expansion

pos indef sg u nom	lik: 270	komp nom	likare: 0
pos indef sg u gen	liks: 0	komp gen	likares: 0
pos indef sg n nom	likt: 321	super indef nom	likast: 0
pos indef sg n gen	likts: 0	super indef gen	likasts: 0
pos indef pl nom	lika: 72	super def no_masc nom	likaste: 0
pos indef pl gen	likas: 0	super def no_masc gen	likastes: 0
pos def sg no_masc nom	lika: 72	super def masc nom	likaste: 0
pos def sg no_masc gen	likas: 0	super def masc gen	likastes: 0
pos def sg masc nom	like: 0	c	lik: 270
pos def sg masc gen	likes: 0	c	lik-: 0
pos def pl nom	lika: 72	sms	lik-: 0
pos def pl gen	likas: 0		

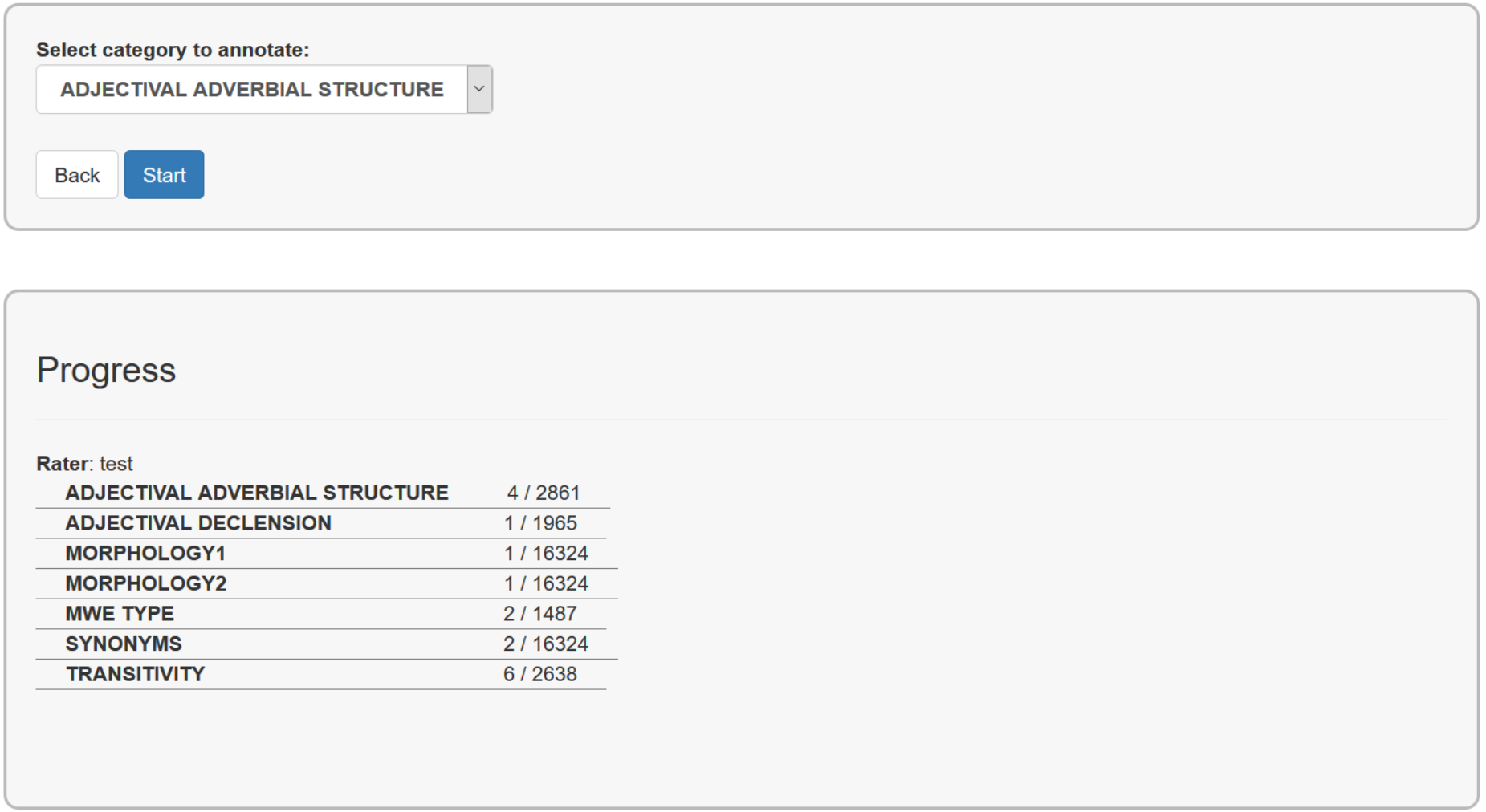
Automatic paradigm expansion

pres ind aktiv	**går bra: 0**
pres ind s-form	**gås bra: 0**
pres konj aktiv	**gånge bra: 0**
pres konj s-form	**gånges bra: 0**
pret ind aktiv	**gick bra: 0**
pret ind s-form	**gicks bra: 0**
pret konj aktiv	**ginge bra: 0**
pret konj s-form	**ginges bra: 0**
imper	**gå bra: 0**
inf aktiv	**gå bra: 0**
inf s-form	**gås bra: 0**
sup aktiv	**gått bra: 0**
sup s-form	**gåtts bra: 0**
pres_part nom	**gående bra: 0**
pres_part gen	**gåendes bra: 0**

Piloting results

Category	IAA 1	IAA 2
nominal declension (6)	0.85	0.80
nominal gender (7)	0.82	0.73
nominal type (5)		0.20
verbal conjugation (14)	0.82	0.94
adjectival declension (2)	0.49	
adjectival adverbial structure (1)	0.39	
morphology 1 (3)		0.48
Overall κ	0.73	0.60

The tool v.2



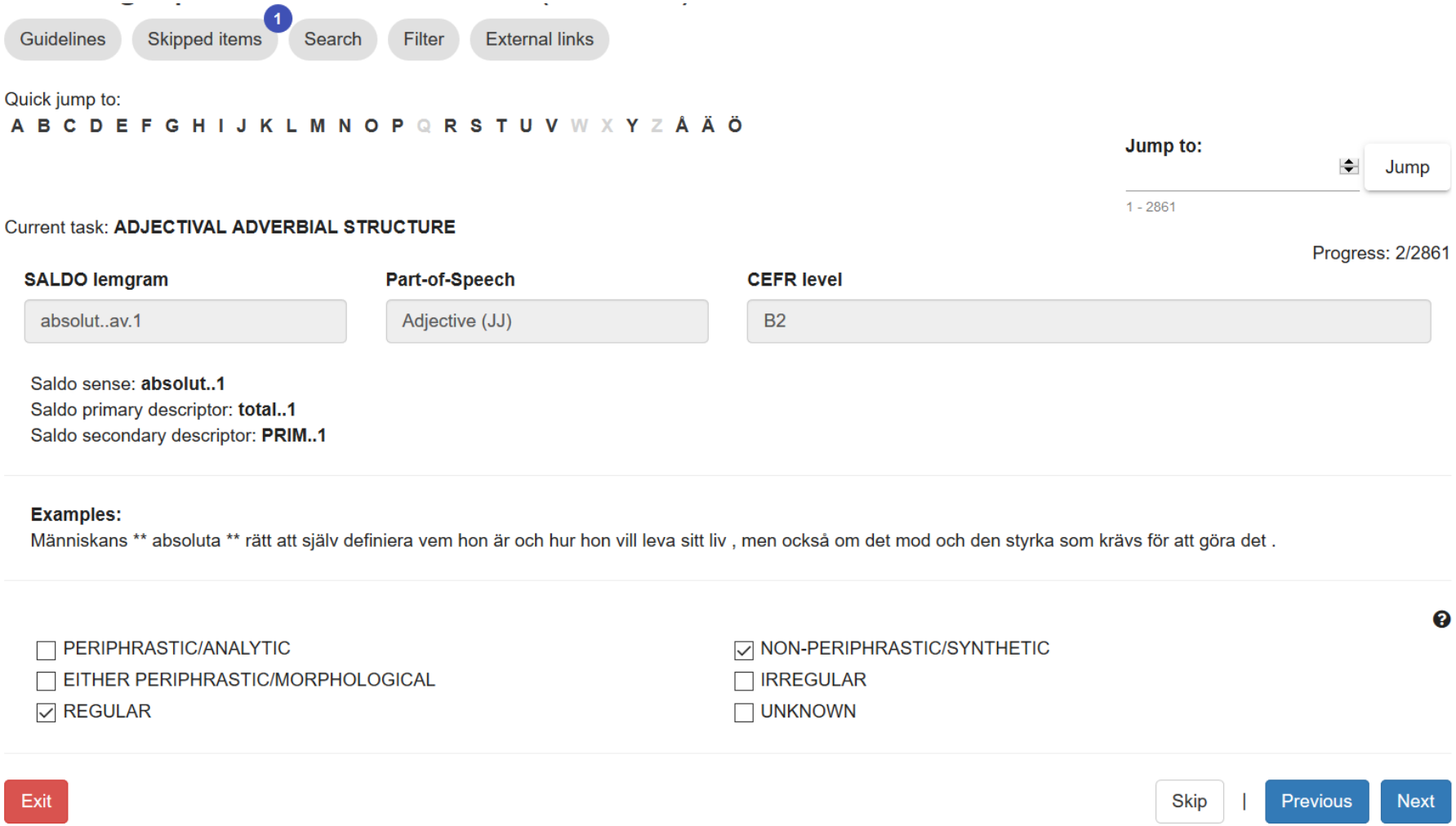
SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

The tool v.2



SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

Demo at NoDaLiDa

David Alfter, Therese Lindström Tiedemann and Elena Volodina.
LEGATO: A flexible lexicographic annotation tool

SPRÅKBANKENTEXT

UNIVERSITY OF HELSINKI

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF GOTHENBURG

References

- Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016, November). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University* (pp. 17-18).
- François, T., Volodina, E., Pilán, I., & Tack, A. (2016, May). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 213-219).
- Volodina, E., Pilán, I., Eide, S. R., & Heidarsson, H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for Computer-Assisted Language Learning* (pp. 128-144).
- Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G., & Sandell, M. (2016). Swell on the rise: Swedish learner language corpus for european reference level studies. *arXiv preprint arXiv:1604.06583*.
- Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016, November). SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016* (No. 130, pp. 76-84). Linköping University Electronic Press.