Workflows for Developing Finnish–German Shallow RBMT... in FinMT, Helsinki, 2017-11-01
https://github.com/flammie/apertium-fin-deu/

Tommi A Pirinen ‹tommi.antero.pirinen@uni-hamburg.de ›

HZSK.de, de.CLARIN.eu, etc.

October 27, 2017

Introduction

- A professional comp.linguist moving to Germany with "no prior knowledge" of the language
- A field/documentary linguistics-heavy North Eurasian language project (SIL FLeX!)
- Prior experience of FSAs, RBMT, SMT...
- ☞RBMT+FSA+language learning+language documentation = cool new workflow for creating: corpora + mono- and bilingual lexicons, chunkers and more**!!**

Language documentation workflow

Field linguists / documentary linguists are used to SIL tools like FLeX (toolbox) with workflows like:

1. split words into morphs
2. tag morphs
3. translate lemmas
4. repeat

Results in a dictionary, annotated corpus and some kind of "morphological analyser".

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkaisi meksikolaisilla stereotypioilla ratsastaneen mainoksen , joka suututti jopa suurlähettiläänn – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkaisi viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalleen , mutta vain muutaman päivän jälkeen mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolaisilla stereotypioilla ratsastaneen mainoksen , joka suututti jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkaisi viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalleen , mutta vain muutaman päivän jälkeen mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolaisilla stereotypioilla ratsastaneen mainoksen , joka suututt–i jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalleen , mutta vain muutaman päivän jälkeen mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–i–lla stereotypioilla ratsastaneen mainoksen , joka suututt–i jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalleen , mutta vain muutaman päivän jälkeen mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–i–lla stereotypio–i–lla ratsastaneen mainoksen , joka suututt–i jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalleen , mutta vain muutaman päivän jälkeen mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainoksen , joka suututt–i jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalleen , mutta vain muutaman päivän jälkeen mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainoksen , joka suututt–i jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutaman päivän jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilään – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viikon lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutaman päivän jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistettu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viiko–n lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutama–n päivä–n jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua siitä .. Teleyhtiö Elisa julkais–i viime viiko–n lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutama–n päivä–n jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua sii–tä .. Teleyhtiö Elisa julkais–i viime viiko–n lopussa jatkoa suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutama–n päivä–n jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua sii–tä .. Teleyhtiö Elisa julkais–i viime viiko–n lopu–ssa jatkoa suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutama–n päivä–n jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua sii–tä .. Teleyhtiö Elisa julkais–i viime viiko–n lopu–ssa jatko–a suositulle Hintasaarnaaja - mainossarjalle–en , mutta vain muutama–n päivä–n jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua sii–tä .. Teleyhtiö Elisa julkais–i viime viiko–n lopu–ssa jatko–a suositu–lle Hintasaarnaaja - mainossarja–lle–en , mutta vain muutama–n päivä–n jälke–en mainos poistettiin kaikkialta .

Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua sii–tä .. Teleyhtiö Elisa julkais–i viime viiko–n lopu–ssa jatko–a suositu–lle Hintasaarnaaja - mainossarja–lle–en , mutta vain muutama–n päivä–n jälke–en mainos poistet–tiin kaikkialta .

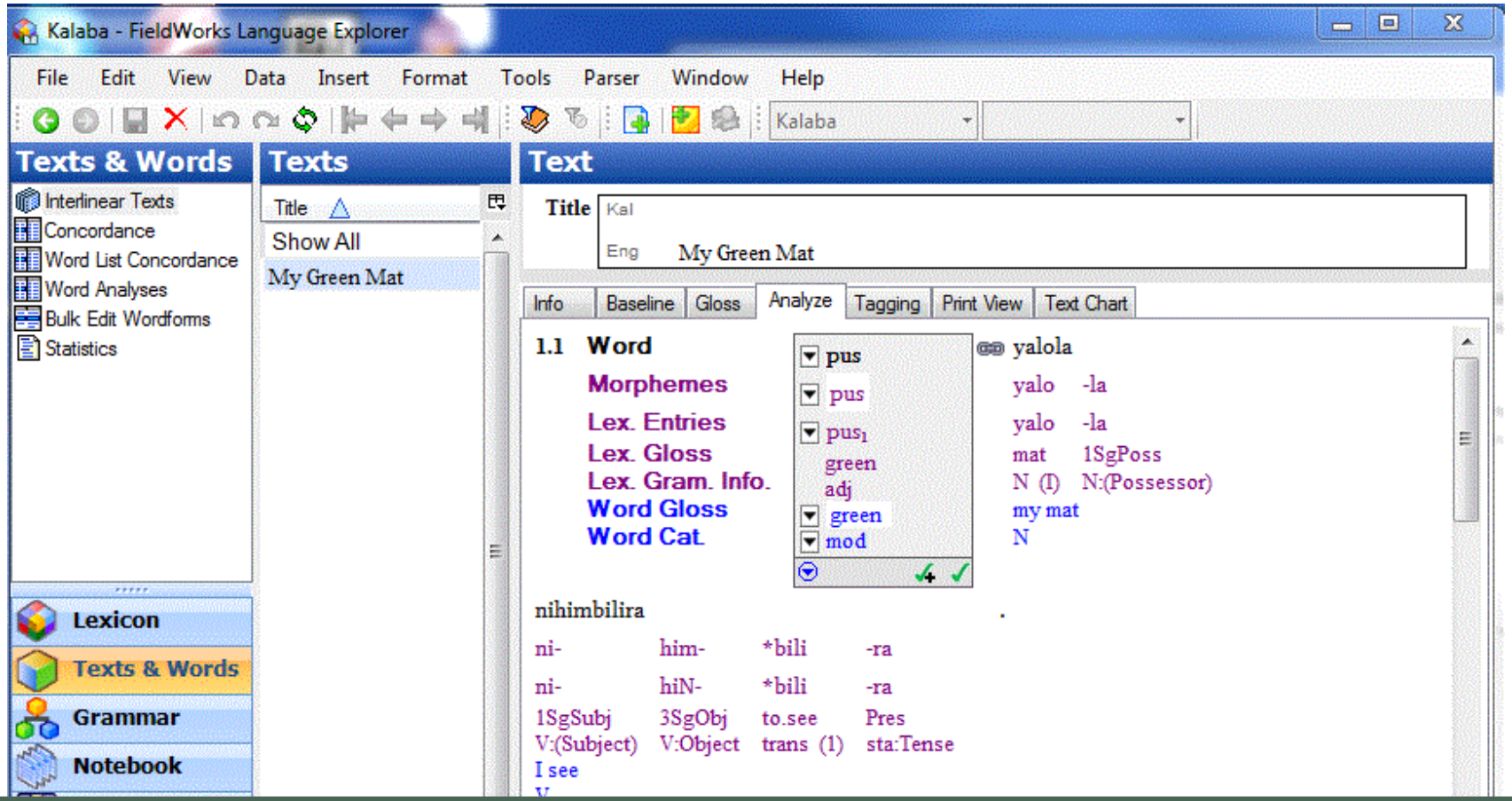
Demo FLeX here (Finnish, simplified)

https://www.hs.fi/nyt/art-2000005425359.html Elisa julkais–i meksikolais–illa stereotypio–illa ratsasta–ne–en mainokse–n , joka suututt–i jopa suurlähettilää–n – Nyt mainos on poistet–tu , eikä yhtiö halua enää edes puhua sii–tä .. Teleyhtiö Elisa julkais–i viime viiko–n lopu–ssa jatko–a suositu–lle Hintasaarnaaja - mainossarja–lle–en , mutta vain muutama–n päivä–n jälke–en mainos poistet–tiin kaikkialta . ... after only 15 minutes of clicking and stuff you get like, past tense, 10 case allomorphs. passive voice, and a dozen of lemma allomorphs

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

hzsk

Real FLeX



RBMT workflow (apertium style)

1. (try to) Translate a text
2. Add OOVs to source language dictionary, repeat 1
3. Add missing word-translations to bilingual dictionary, repeat 1
4. Add ungenerateable words to target language dictionary, repeat 1
5. Mangle mismatching lexical grammar
6. (Advanced RBMT not relevant to this presentation, magic)

Results in two morphological analyser/generators, a bilingual dictionary and a shallow comparative grammar!

RBMT workflow demo

Am S-Bahnhof Berliner Tor hat es gleich drei Einsätze der Bundespolizei gegeben. Donnerstagabend war dort ein stark angetrunkener Mann auf die Gleise gefallen und musste von den Beamten gerettet werden. Freitag um halb acht lief ein ebenfalls angetrunkener Mann auf den Gleisen Richtung Hauptbahnhof. Und gegen zehn warf ein Mann den Deckel eines Mülleimers auf die Gleise. In allen drei Fällen musste die Strecke kurzzeitig gesperrt werden.

RBMT workflow demo

Päälle *S-asema berliiniläinen portilla on se samoin kolme käytöt liittovaltionpoliisin annettu. torstai-ilta oli siellä yksi vahvasti *angetrunkener mies @Gleis<n><pl><ine> miellyttää ja täytyi virkailijoista pelastetaan. perjantai jotta puolittainen kahdeksan saapui myös *angetrunkener mies @Gleis<n><pl><nom> suunnassa @Hauptbahnhof<n><sg><nom>. Ja vastaan kymmenen pudotti mies @Deckel<n><sg><gen> @Mülleimer<n><sg><gen> @Gleis<n><pl><ine>. kaiken kolme tapauksien/tapausten täytyi reitti lyhytaikainen estetään.

RBMT workflow demo

~~Päälle~~ @S-Bahnhof<n><sg><nom> berliiniläiseltä portilla on se samoin kolme käytöt liittovaltionpoliisin annettu. torstai-ilta oli siellä yksi vahvasti @angetrunken<adj><pos><sg><nom> mies @Gleis<n><pl><ine> miellyttää ja täytyi virkailijoista pelastetaan. perjantai jotta puolittainen kahdeksan saapui myös @angetrunken<adj><pos><sg><nom> mies @Gleis<n><pl><nom> suunnassa @Hauptbahnhof<n><sg><nom>. Ja vastaan kymmenen pudotti mies @Deckel<n><sg><gen> @Mülleimer<n><sg><gen> @Gleis<n><pl><ine>. kaiken kolme tapauksien/tapausten täytyi reitti lyhytaikainen estetään.

RBMT workflow demo

#lähijuna-asema<n><sg><nom> berliiniläiseltä portilla on se samoin kolme käytöt liittovaltionpoliisin annettu. torstai-ilta oli siellä yksi vahvasti humalainen mies raiteissa miellyttää ja täytyi virkailijoista pelastetaan. perjantai jotta puolittainen kahdeksan saapui myös humalainen mies raiteet suunnassa #päärautatieasema<n><sg><nom>. Ja vastaan kymmenen pudotti mies kannen roskiksen raiteissa.

RBMT workflow demo

lähijuna-asema berliiniläiseltä portilla on se samoin kolme käytöt liittovaltionpoliisin annettu. torstai-ilta oli siellä yksi vahvasti humalainen mies raiteissa miellyttää ja täytyi virkailijoista pelastetaan. perjantai jotta puolittainen kahdeksan saapui myös humalainen mies raiteet suunnassa päärautatieasema. Ja vastaan kymmenen pudotti mies kannen roskiksen raiteissa. kaiken kolme tapauksien/tapausten täytyi reitti lyhytaikainen estetään.

Experimental "results"

- 1.5 years after starting at week-daily average effort of 15 minutes
- approx. 10,000 translation pairs, few hundreds of new lexemes for analysers, few dozens of rules for WSD, chunking, lex.sel., etc.
- Post-edition WER is about 40, BLEU probably 8 or so,
- ... but it's easy to read and understand!
- For realistic projects, e.g. within Uralic group, it should be perfectly plausible

Let's go: `http://39476.s.time4vps.cloud`