Lexical Resources in GF

Krasimir Angelov

University of Gothenburg

July 15, 2015

Some History

- 2008 OALD imported
  (Björn Bringert)
- 2010 Further development for wide coverage parsing in English
  (Krasimir Angelov)
- 2012 Translation to Swedish, Finnish, Hindi, Urdu, Bulgarian
  (Aarne Ranta, Shafqat Virk, Krasimir Angelov)
- 2013 First Mobile Translator
  (Björn Bringert, Krasimir Angelov)
- .... Many more languages added

English Lexicon

- Nouns, Verbs, Adjectives, Adverbs
  - Oxford Advanced Learners Dictionary
  - Princeton WordNet
  - Spelling variants (British/American/Others)
  - Harmonized with RGL
- Prepositions
  - PennTreebank
  - Wikipedia
- Verb Frames
  - PennTreebank
  - VerbNet (TODO)
- Phrasal Verbs
  - Web Sites for Learning English

English Lexicon

Example:

```
lin house_N = mkN "house" "houses";
lin play_V = mkV "play";
lin beautiful_A = compoundA (mkA "beautiful");
lin behind_Adv = mkAdv "behind";
lin instead_of_Prep = mkPrep "instead of";
lin theatre_N  = variants {mkN "theatre";
                           mkN "theater"};
lin maharaja_N = variants {mkN "maharaja";
                           mkN "maharajah"};
```

Verb Frames

- Currently a limited inventory of verb frames from OALD and PennTreebank

```
lin make_V = IrregEng.make_V;
lin make_V2 = mkV2 (IrregEng.make_V);
lin make_V2A = mkV2A (IrregEng.make_V) noPrep;
lin make_V2V = mkV2V (IrregEng.make_V) noPrep noPrep;
```

- VerbNet has a better inventory which should be incorporated. This would also require extensions in the RGL

Multiword Units

- There are a number of multiword units:

```
lin cod_liver_oil_N = mkN "cod-liver oil" ;
```

- These are all inherited and there is no clear criteria about which units should be in the lexicon.

Translations

- Free Electronic Dictionaries (Bulgarian, Swedish)
- WordNet (Finnish)
- Universal WordNet (Bulgarian)
- Apertium (Bulgarian, Others?)
- Google Translate (Bulgarian, Swedish)
- Phrase Tables (Bulgarian)
- PannLex (Thai)
- Manual Translation (Bulgarian, Chinese)
- Wiktionary (Most Other Languages)

Sense Splits

- Sense Ambiguities in English

	English	Swedish
letter_1_N	letter	brev
letter_2_N	letter	bokstav

- Gender Ambiguities in English

	English	Bulgarian	German
teacherMasc_N	teacher	učitel	Lehrer
teacherFem_N	teacher	učitelka	Lehrerin

Morphology

- Smart Paradigms
- IrregXXX modules
- Free Morphological Lexicons
  (OALD, Open Office, SALDO, KOTUS)

Validation

There are still many errors in the dictionaries. English, Swedish and Bulgarian seems to be in the best shape.

- Go Through the Word List in Frequency Order
- Use Your Vacation to Test the Translator

GF Lexicon vs WordNet

GF Lexicon

- Mostly one sense per word
- Focus on the primary sense
- Many sense confusions

WordNet

- No morphology
- Coarse POS tags
- Not focused on translation

Ongoing and Past Work on Integration

Past

- Shafqat Virk, K.V.S. Prasad, Aarne Ranta, Krasimir Angelov. Developing an interlingual translation lexicon using WordNets and Grammatical Framework.

Ongoing

- Selective Translation Choice from WordNet

A New Statistical Model

- The current model is trained on the English PennTreebank
- With more split senses we will need something else:
  - Princeton WordNet has some sense frequency information
  - This can be complemented by using the EM algorithm.
    Example:

	English	Swedish	Bulgarian	German
letter_1_N	letter	brev	pismo	Brief
letter_2_N	letter	bokstav	bukva	Buchstabe
teacherMasc_N	teacher	lärare	učitel	Lehrer
teacherFem_N	teacher	lärare	učitelka	Lehrerin