

## **Automating SFL Annotation: what we can do, what is left to do**

**Mick O'Donnell**  
Universidad Autónoma de Madrid

### **1. Building Automatic Annotation Tools for SFL**

1. Quick development of annotated corpora depends on automatic annotation tools
2. Parsers depending on hand-written grammars/lexicons take too long to develop.
3. Statistical parsers require an existing treebank of analyses to drive them.
4. We lack systemic-oriented treebanks.
5. Thus, we cannot produce statistical parsers for SFL annotation at present.

## 2. Reusing existing Tools

- However, sizable annotated corpora exist for PSG style analyses:
  - Penn Treebank, etc.
- Fairly reliable Statistical parsers exist for English, and some other languages (German, Chinese, Arabic?)
  - Collins Parser,
  - Stanford Parser

## 2. Stanford Parser

```
(ROOT
(S
(NP
(NP (NNP President) (NNP George) (NNP W.) (NNP Bush))
(PP (IN on) (NP (NNP Saturday))))
(VP (VBD defended)
(NP (PRP$ his)
(NP (QP ($ $) (CD US350) (CD billion))
(NN tax))
(NN cut) (NN package))
(PP (IN against)
(NP (NN opposition) (NNS accusations)))
(SBAR (IN that)
(S
(NP (PRP it))
(ADVP (RB unfairly))
(VP (VBZ benefits)
(NP (DT the) (JJ rich))))))
(. )))
```

President George W.  
Bush on Saturday  
defended his \$US350  
billion tax cut package  
against opposition  
accusations that it  
unfairly benefits the  
rich.

## 2. Stanford Parser

(ROOT  
(S  
(NP (NN Recht) (NE groß))  
(VAFIN ist)  
(NP (ART die) (NN Lücke)  
(PP (APPRART am) (NN Vorplatz)  
(CNP  
(NP (ART des) (NN Darmstadtiums))  
(KON und)  
(NP (ART der) (NN Einmündung)  
(NP (ART der) (NN Alexanderstraße.))))))

Recht groß ist die  
Lücke am Vorplatz des  
Darmstadtiums und  
der Einmündung der  
Alexanderstraße.

## 3. Questions for SFLers

1. Can we use these parsers as a starting point to building our own SFL-annotated corpora?
2. And can we then use these corpora to produce Statistical parsers for SFL without need of the PSG parsers?

#### 4. Using Existing Parsers to do SFL

- UAM CorpusTool makes use of the Stanford Parser.

##### **Current Status:**

- Stanford parse tree used for automatic segmentation into clauses or NPs.
- Mistakes are made, but the annotation interface allows the user to correct mistakes.
- English incorporated.
- Debugging German
- Chinese, Arabic to do.
- Looking for equivalent Spanish parser or treebank

#### 4. Using Existing Parsers to do SFL

- DEMO

#### 4. Using Existing Parsers to do SFL

##### **Future Development:**

- Stanford parse tree used as skeleton for deriving SFL analyses of the sentences:
  - Stanford NP-VP analysis mapped to a Mood analysis: Subj-Pred-Obj etc.
  - This Mood analysis used to derive Transitivity analyses (Actor, etc.), using other resources such as semantic lexicons.
  - Theme analysis also derived from Mood analysis.
  - Map different parsers to same structure

#### 4. Using Existing Parsers to do SFL

- My Goal is to develop within CorpusTool a means for users to write rules which specify how one representation should be mapped onto another
- E.g.  
Identify experiential-theme if “^Subject”

#### 4. Using Existing Parsers to do SFL

- My Goal is to develop within CorpusTool a means for users to write rules which specify how one representation should be mapped onto another

- E.g.

**Identify experiential-theme** if "**^Subject**"

Create a segment in the current layer  
and assign it the feature:  
"experiential-theme"

#### 4. Using Existing Parsers to do SFL

- My Goal is to develop within CorpusTool a means for users to write rules which specify how one representation should be mapped onto another

- E.g.

Identify experiential-theme **if** "**^Subject**"

Create the segment for any text which  
is tagged as Subject at another layer,  
And that subject is the first constituent  
in its parent unit.

#### 4. Using Existing Parsers to do SFL: Limits

- SFL analyses contain more information than PSG analyses (closer to semantics)
- So, mapping from PSG to SFG cannot be done without adding information from somewhere.
- Partially, this can be done by making reference to a semantically rich lexicon, providing, e.g., possible process types of verbs, conceptual nature of nouns, etc.
- Additionally, we need information about collocation of Systemic categories: e.g. mental processes (generally) require a human Sensor, etc.

#### 4. Using Existing Parsers to do SFL: Limits

- Lastly, human annotators can provide the extra information needed, letting the computer make the most likely mapping, and allowing the human to correct mistakes.
- Or is it better for the computer to prompt the user whenever it cannot decide?
- My biggest issues in incorporating automatic annotation in CorpusTool relates to HOW the intelligent software and the human work together:
  - Auto-analysis with post editing (e.g., clause segmentation)
- Vs.
  - Human checking at each step (e.g., the Autocoder interface)

#### 4. Using Existing Parsers to do SFL: **Conclusions**

##### **Conclusions**

- We can use non-SFL parsers to help construct SFL annotated corpora.
- However, this needs to include human intervention (post-editing or 'checkpoint' style).
- Additional resources required support semantically rich classification:
  - Semantic lexicons
  - Collocation patterns

#### 5. Towards an SFL Statistical Parser?

- *Can we use an SFL-annotated corpus to produce a statistical parser which produces SFL analyses directly, without recourse to an intermediate PSG analysis?*
- **Problem:** the more detail we have in the analysis, the more examples we need in our corpus.
  - Penn treebank: small set of classes for words and phrases (e.g., just NP for an NP).
- For a SFL analysis, we use a lot more labelling.
  - 2 kinds of labelling rather than one: function and class.
  - 3+ layers of grammatical analysis
  - Semantically oriented classes and functions
- **THUS: far larger corpus required!!!!**

## 5. Towards an SFL Statistical Parser?

- **Problem 2:** where a program needs to resolve ambiguity, the more alternatives there are, the more chance it has of getting it wrong
  - -> lower rate of accuracy.
- **Problem 3:** PSG categories relate strongly to structural evidence.
- SFL categories also relate to how the unit will function in a wider context (and context may not be available).

## 5. Towards an SFL Statistical Parser: **Conclusions**

### **Conclusions:**

- We could produce an SFL parser based on an SFL corpus, but we would need a larger corpus than used by existing statistical parsers.
- Our rate of accuracy will be lower, but may still be usable.
- One never gets anywhere by saying:
  - “It is too hard. Why bother!”