

# Annoterade satser och platta frasstrukturer för svenska



---

Johnny Bigert  
Ola Knutsson  
Viggo Kann  
Jonas Sjöbergh



# Bygga av svensk trädbank

---

- SUC - korrekt, tillgängligt, balanserat
- Majoritetsröstning med befintliga verktyg för frasstrukturer - halvautomatik
- Manuell annotering / manuell rättning?



# Utvärdering

---

- Manuell utvärdering:  
små ändringar - mycket arbete
- Automatisk utvärdering - individuella delars inverkan på prestanda



# Detektion av svåra stavfel

---

Hybridmetod:

- Statistisk/automatisk del
- Frasidentifiering - manuella regler

Exempel på feltyper:

- Stavfel resulterande i befintliga ord
- Saknade ord



# Utvärdering av hybrid

---

Vi påverkas av fel i de olika modulerna:

- Tokenisering, taggning, frasidentifiering
- Hur bra kan metoden bli med perfekt data (maxprestanda)?



# Utvärdering av hybrid

---

## Robusthet:

- Vad blir effekten om ingen hänsyn tas till kongruens?
- Taggningsfel, fritext innehåller fel
- Svåra undantag för NP-kongruens i svensk grammatik
- Sämre prestanda med kongruens i NP



# Diskussion

---

- Manuell annotering / manuell rättning?  
On-line, Annotate (DFKI)?
- SUC?
- Detaljnivå på gemensamt format?
- Startdatum?