Glossing a multi-purpose corpus

Ellen Ormel & Onno Crasborn

Radboud Universiteit Nijmegen
www.ru.nl/sign-lang

Context

Corpus NGT: 72 hours, 2375 clips. Initial release in 2008: rough glosses for 10 hours. Open access through MPI corpus browser corpus1.mpi.nl and SurfMedia streaming service for higher education.
www.ru.nl/corpusngtuk

SignSpeak: EU project 2009-2012. Goal: automatic translation of sign videos to written text. The primary sources are Corpus NGT movies.
www.signspeak.eu

Automatic sign recognition and translation: variable visual input > ID-gloss > contextually appropriate sentence-level translation
SignSpeak

- Computer is stupid. Computer needs practice.
- Type/token ratio of min. 1/15
- Max. no. of singletons should be 40% of the lexicon

General problems

- Glosses are often not a representation of the form of the language under investigation, but labels in the spoken/written language used by researchers

\[
\text{Russian} \\
\text{My s Marko poexa-l-i autobus-om v Peredelkino.} \\
\text{1PL COM Marko go-PST-PL bus-INS All Peredelkino} \\
\text{we with Marko go-PST-PL bus-by to Peredelkino} \\
\text{‘Marko and I went to Peredelkino by bus.’}
\]

- ID-glosses: references to a lexicon
  - No usable electronic lexicon for NGT
- No conventions on how to represent the morphology of the language if we would use HamNoSys for glossing the source text (no simultaneous interlinear glossing)
Our problems

• No lexicon

• Initial glossing of the Corpus NGT (2007-2008, ±10hrs):
  quick and dirty: let’s try to label as many events in the video as we can
  intuitive: let the 6 different deaf annotators decide what is a good label

• Results
  • many translation variants used for same form
  • no labeling of different (regional or other) variants

• No phonetic/phonological annotation

Steps in improving Corpus NGT glosses

1. Improve annotation manual: be more explicit, more examples
2. Be stricter in sticking to conventions
3. Construct elementary sign list (dictionary?)
   • phonological form
   • possible meanings in different contexts
   • variant markers –A –B –C (not distinguishing regional from other variation)
   • different glosses for manual homonyms with different mouthings
4. New conventions
   • meaning of lexical item in this context can be added on child tier
   • improved annotations of numerical constructions
   • separate tier for mouth activity available during glossing
Suffixes -A and -B indicate there are different forms referred to by this gloss

1. DOCTOR-B
2. BATTERY

Homonym signs receive two different glosses
Phonological description in the lexicon

The lowercase words are possible translations of the sign in English

TRY-A
TRY-B

try, attempt

CONTENT
happy, content, satisfied, pleased, comfortable, at ease

Signs sometimes have a large set of possible translations in Dutch or English; the best equivalent can be added after the gloss, e.g. CONTENT>comfortable
RELATIONSHIP

Keywords in dictionary / possible translation variants in various contexts:
  relationship
  husband
  wife
  married
  wedding
  etc.

Current work

• Continue with creation of lexicon
• Apply these ID-glosses to the existing corpus
  >95,000 glosses, L+R combined
• Start with using lexicon for new glosses

• A-B-C variants will decrease the average type/token ratio
• One-form-one-gloss will increase the average type/token ratio: no longer different glosses for the same manual form (modulo the manual homonyms)
Concluding remarks

• Creation of a lexicon has proven to be crucial to ensure promote glossing consistency
  • form variants without obvious semantic difference
  • either lemmata or full word forms are OK
• Annotators observe a lot of interesting ‘trivia’ that you don’t want to lose (but shouldn’t impact the gloss either)
  • new tiers per signer: mouthing, meaning, repetition
Thank you

www.ru.nl/sign-lang
www.ru.nl/corpusngt
www.signspeak.eu