Corpus informed lexicography: a decade of exploration

Rachel McKee & David McKee

‘Sign Language Corpora: Linguistic Issues’ Workshop
DCAL, University College London, July 25th 2009
Reliability depends on the kind of evidence underpinning a dictionary

- **Citations (from texts)** are usual form of evidence for creating conventional dictionary entries

- **Introspection & informant testing** - common source of evidence for unwritten languages: a **subjective** basis
Editor Graeme Kennedy: non-signer, with lexicography & corpus expertise

\[ \therefore \text{DNZSL should have empirical, descriptive basis} \]

- ‘Concept net’ design - capture topics and concepts common to most languages

- Video-recorded NZSL discussion groups on these topics > a corpus for describing lexicon
  - 4,500 signs (incl. variants) in dictionary
Zipf’s law: words are used (distributed) with different frequencies

- A few words account for a high % of any text.
- eg, 100 English words = 45% of 100 million words in British National Corpus

Need to reliably identify most frequent signs for concise dictionary
Wellington Corpus of NZSL compiled

- 40 hours of tape: dictionary + other recordings
- 80 Deaf people, age 18-60
- Range of topics; dialogue & monologue
- Gloss transcription (not annotated)
- Took 1 year, 1 person full-time
- 100,000 running signs
KNOW FAMOUS HORSE AUSTRALIA 1 MAN FROM HERE HORSE-TROTTING NAME fs-SHANUE fs-DYE PRAM KNOW MELBOURNE MELBOURNE CUP WIN FIRST SCL-1-horse-leads FIRST 3 YEAR PAST 3 YEAR PAST IX-3 POS1 FRIEND IX-3 IX-1 LONG-TIME-AGO SMALL-CHILD IX-3 FATHER IX-3 fs-WAS POS1 FRIEND IX-3 IX-1 IX-3 NMS-nod IX-3 IX-3 BORN IX-3 IX-3 FATHER SISTER IX-3 FAMILY AREA OLD GOOD TOGETHER GOOD EACH-OTHER IX-3 IX-3 fs-SON IX-3 GROW-UP MELBOURNE FAMOUS HORSE CUP IX-3 GOOD IX-3 NOW GOOD AREA AUSTRALIA HORSE COMPETE GOOD POS3 FATHER POS3 SISTER IX-3 IX-1 LEARNER-LICENCE KISS NOW IX-3 LEARNER-LICENCE IX-1 IX-3 TEACH-me IX-1 LEARNER-LICENCE WELL WHEN GROW-UP LATE 12 13 DEAF LIVE MOVE NOW HOUSE MAORI WITH IX-1 CLASS IX-1 IX-3 IX-1 IX-3 GOOD IX-3 CHILDREN 8 CHILDREN 8 MAORI 8 CHILDREN WOW 1 FROM 1 DEAF IX-3 FATHER-MOTHER GOOD KIND-TC GIVE-me FOOD PROVIDE FOOD IX-1 EAT POS3 MAORI WAY BREAD PCL-B-heaps-of-bread CUT-BREAD […] IX-1 SMALL-CHILD IX-3 GET-AWAY BIG-KID-2h COMPETE IX-1 ONCE SEE SOMEONE TENNIS COURT SOMEONE CATCH CRAB BOX DCL-BB-box POSSUM fs-OPOSSUM POSSUM
Used Wordsmith (concordance tool) to analyse distribution of lexical items for purposes of:

- **Concise Dictionary** content - high freq vocab
- **Teaching** – most ‘useful’ vocab to learn/teach
  - *How many & which signs are needed for everyday communication in NZSL?*

Findings about types

• In 100,000 tokens (running signs)
  – 7,222 lexical types (distinct glosses)
  – Including 1,079 FS types (full & single-letter forms)
    • 2,554 tokens FS = 2.5% of corpus
  – Polysemy & non-frozen lexicon reduce the number of lexical types in a SL corpus
    • But … large relative to the number of signs in most SL dictionaries
<table>
<thead>
<tr>
<th>Percent of Engl/NZSL corpus covered</th>
<th>by number of English Word types</th>
<th>by number of NZSL Sign types</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>10 -15</td>
<td>11 (= 20%)</td>
</tr>
<tr>
<td>50 %</td>
<td>50 -100</td>
<td>116</td>
</tr>
<tr>
<td>70 %</td>
<td>--</td>
<td>343</td>
</tr>
<tr>
<td>80%</td>
<td>1,000 -1,500</td>
<td>665</td>
</tr>
</tbody>
</table>

Potentially – a learner who knows the most frequent 665 signs can access 80% of vocab in NZSL discourse - compared to a 1,000-1500 ‘basic vocab’ for English
# Top 12 signs (20% of corpus)

<table>
<thead>
<tr>
<th></th>
<th>Sign Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>IX-1 (I, me, we, us)</td>
<td>6,720</td>
</tr>
<tr>
<td>2.</td>
<td>IX-3 (he, she, it)</td>
<td>3,648</td>
</tr>
<tr>
<td>3.</td>
<td>GOOD</td>
<td>1,462</td>
</tr>
<tr>
<td>4.</td>
<td>DEAF</td>
<td>1,419</td>
</tr>
<tr>
<td>5.</td>
<td>IX-2 (you)</td>
<td>1,153</td>
</tr>
<tr>
<td>6.</td>
<td>POS-1 (my, mine)</td>
<td>1,095</td>
</tr>
<tr>
<td>7.</td>
<td>IX-loc (there)</td>
<td>914</td>
</tr>
<tr>
<td>8.</td>
<td>ONE</td>
<td>677</td>
</tr>
<tr>
<td>9.</td>
<td>SAME</td>
<td>669</td>
</tr>
<tr>
<td>10.</td>
<td>SCHOOL</td>
<td>658</td>
</tr>
<tr>
<td>11.</td>
<td>YES</td>
<td>643</td>
</tr>
<tr>
<td>12.</td>
<td>SIGN</td>
<td>626</td>
</tr>
</tbody>
</table>
• 194 high freq concepts *not* in English top 350
• Deaf culture themes
  – **Communication**: DEAF, HEARING, SIGN, ORAL, EXPLAIN, PAST-MY-EYES, COMMUNICATE, SIGN-CHAT
  – **School**: TEACHER, KELSTON, BOARDER, CLASS
  – **Deaf community**: CLUB, SPORT, ASSOC’N, CL- gather
• **Verbinesss**: high % of top 350 are verbs
  GO, HAVE, SAY, WORK, THINK, SEE, KNOW, WANT, LOOK, FEEL
• **English influence**: 2.5% of tokens are fingerspelling: 14.9% of all types - but most are low freq items
Limitations of WCNZSL

• **Size & composition**
  – 100,000 signs large for a sign corpus (cf. Morford & Macfarlane 2003: 4,111 signs), but still small scale
  – Representativeness of topics, genres, speakers?

• **Consistency of glossing** (not 100% ID glosses)

• **Not video linked** - hard to retrieve original source

• **‘Bare’ manual lexemes only**
  – No annotation of other features

• **Synchronic** – sample of NZSL at one point in time
Freelex is an open source database application for dictionary making

designed by Dave Moskovitz

Download software

http://www.matapuna.org/

Online Dictionary of NZSL (in progress)

http://nzsl.vuw.ac.nz/dnzsl/freelex/freelex

Database links to a corpus search function
Sample of 150 fluent NZ signers – stratified by region, age group, ethnicity, gender
  – approx. 100 hours of conversation, interview

So far, 81 excerpts of 1-2 mins each transcribed in ELAN
  – Annotated target features for variation analysis
• Variation text files (from ELAN) = 14,000 signs added to Wellington Corpus.

• Now using this combined corpus in the online dictionary to inform
  – senses & usage, semantic context of signs
  – basis of example sentences
  – word class & collocation information
Moniker
lucky

id #
2701

Variant Number
1

Main Gloses
lucky

Secondary Gloses
Fortunate, fortunately, good luck, luck

Minor Gloses

Example comments

Word classes
- adjective
- interjection
- interrogative
- negator
- noun
- numeral
- phrase
- pronoun
- verb

final.example1
me fail ix-3 get job lucky he

final.example2
me baby born short labour four-hour feel good me lucky

final.example3

final.example4

asset
picture-W30-37.png [delete]
<table>
<thead>
<tr>
<th>Example</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>example1</td>
<td>MOTHER RUBELLA BUT POS1 MOTHER IX-1 NEARLY BLIND DEAF LUCKY MISS OUT</td>
</tr>
<tr>
<td>example1source</td>
<td>WC_Eddie_Bridget_Stirrat_Maureen_Tompson_education</td>
</tr>
<tr>
<td>example2</td>
<td>EDUCATION LUCKY GOOD EDUCATION HELP-me IX-1 GET ACCESS TO INFORMATION AND READ WRITE STUDY</td>
</tr>
<tr>
<td>example2source</td>
<td>WC_Panel_discussion_Akoranga</td>
</tr>
<tr>
<td>example3</td>
<td>REALLY-SUCCEED BEST EDUCATION WRITE INTERESTING LUCKY POS1 MOTHER SCHOOL TEACHER BEHIND PUSH WRITE EVERYDAY HOME WRIT</td>
</tr>
<tr>
<td>example3source</td>
<td>WC_Wayne_Bird_Interview</td>
</tr>
<tr>
<td>example4</td>
<td>IX-1 FAIL IX-3 GET JOB LUCKY IX-3</td>
</tr>
<tr>
<td>example4source</td>
<td>WC_Brent_MacPherson_life_narrative</td>
</tr>
<tr>
<td>example5</td>
<td>IX-1 SHORT fs-L-LABOUR FOUR-HOUR FEEL GOOD IX-1 LUCKY</td>
</tr>
<tr>
<td>example5source</td>
<td>WC_Julie-Anne_Taylor_birth</td>
</tr>
<tr>
<td>example6</td>
<td>PRO1 KNOW-ix POSS-1 CHILDREN HIGH SCHOOL EXPENSIVE UNIFORM-1h LUCKY ns-NAMESIGN-xx PRO3 STILL-2 SAME-throughout UNIFORM FOR YEAR PRO1 PAY NOTHING</td>
</tr>
<tr>
<td>example6source</td>
<td>SV_Leanne_Holland-charee_leanne_uniform_2</td>
</tr>
</tbody>
</table>
There were 48 hits for lucky in 28 files.
Using corpus for entry info

- **Checking senses**
  - ‘FIGHT’ - literal & metaphorical (English) senses?

- **Context, word class**
  - AFFAIR - verb/noun?
  - HOT - only with +human subject?
  - DELEGATE - mainly/only with sport?

- **Mouthing, NMF** (for filming examples)
Benefits of corpus examples

1. Cultural relevance of contexts
   • compare *original dictionary (constructed)* examples given for LUCKY:
     – (Adj) I was **lucky** to win the raffle.
     – (Adv) **Fortunately** we missed the traffic.

     with *NZSL corpus examples*
     – MOTHER RUBELLA BUT POS1 MOTHER IX-1 NEARLY BLIND DEAF **LUCKY** MISS OUT
     – **LUCKY** GOOD EDUCATION HELP-me IX-1 GET ACCESS TO INFORMATION AND READ WRITE STUDY

2. Show word class, collocation & syntax
Creating usage examples from a corpus: Criteria (Atkins & Rundell)

1. Natural & Typical
   word in most usual context, syntax and collocation; not idiosyncratic usage; not mixing registers or varieties

2. Informative
   sentence gives informative context (helps understand sense of word)

3. Intelligible
   contains no words that are more difficult than the headword; clear structure; succinct
• 50% of dictionary headwords not found in corpus:
  – headword/ gloss differences
  – limited size of corpus

• Natural utterances maybe not accessible to learners
  – Complex or fragmented structure
  – Low frequency or complex signs in sentence
  – Meaning is too contextualised (sentence can’t stand alone)
  – Example doesn’t reflect most ‘typical’ meaning

• Re-performing sentences from exact glosses not easy (for making dictionary video clips)

• Pragmatic compromise: corpus informed, but modified, usage examples
References

- Sinclair, J. 1991 Corpus, Concordance, and Collocation. OUP

Contact us: rachel.mckee@vuw.ac.nz, david.mckee@vuw.ac.nz