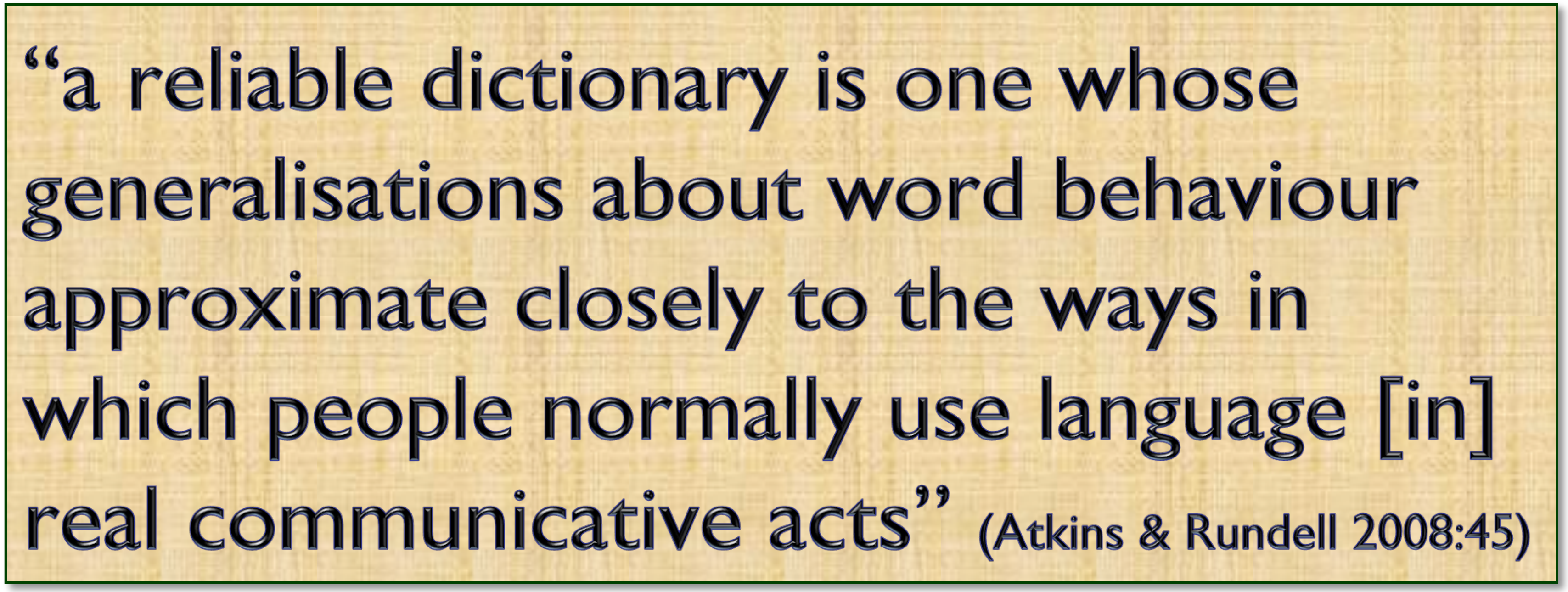
Corpus informed lexicography: a decade of exploration

Rachel McKee & David McKee

‘Sign Language Corpora: Linguistic Issues’ Workshop

DCAL, University College London, July 25th 2009

Victoria
UNIVERSITY OF WELLINGTON
Te Whare Wānanga
o te Ūpoko o te Ika a Māui



"a reliable dictionary is one whose generalisations about word behaviour approximate closely to the ways in which people normally use language [in] real communicative acts" (Atkins & Rundell 2008:45)

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

1992 Dictionary of NZSL

- Editor Graeme Kennedy: non-signer, with lexicography & corpus expertise

  ∴ ***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

1997 – Concise Dictionary
Which 2,000 signs to include?

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

IX-2 KNOW IX-2 FAMOUS HORSE IX-loc AUSTRALIA IX-loc 1 MAN FROM HERE HORSE-TROTTING NAME fs-SHANUE fs-DYE PRAM KNOW IX-2 FAMOUS IX-3 MELBOURNE MELBOURNE CUP 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 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

Distributional analysis

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?*

See: McKee, David & Graeme Kennedy 2006. The Distribution of Signs in New Zealand Sign Language. *Sign Language Studies* 6 (4). 372-390

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

Coverage of corpus by types: English vs NZSL

Percent of Engl/NZSL corpus covered	by number of English Word types	by number of NZSL Sign types
25%	10 -15	**11 (= 20%)**
50 %	50 -100	**116**
70 %	--	**343**
80%	1,000 -1,500	**665**

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)

1.	IX-1 (I, me, we, us)	6,720
2.	IX-3 (he, she, it)	3,648
3.	GOOD	1,462
4.	DEAF	1,419
5.	IX-2 (you)	1,153
6.	POS-1 (my, mine)	1,095
7.	IX-loc (there)	914
8.	ONE	677
9.	SAME	669
10.	SCHOOL	658
11.	YES	643
12.	SIGN	626

Features of NZSL lexicon

- 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

Online Dictionary of NZSL project 2008-2011

- **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**

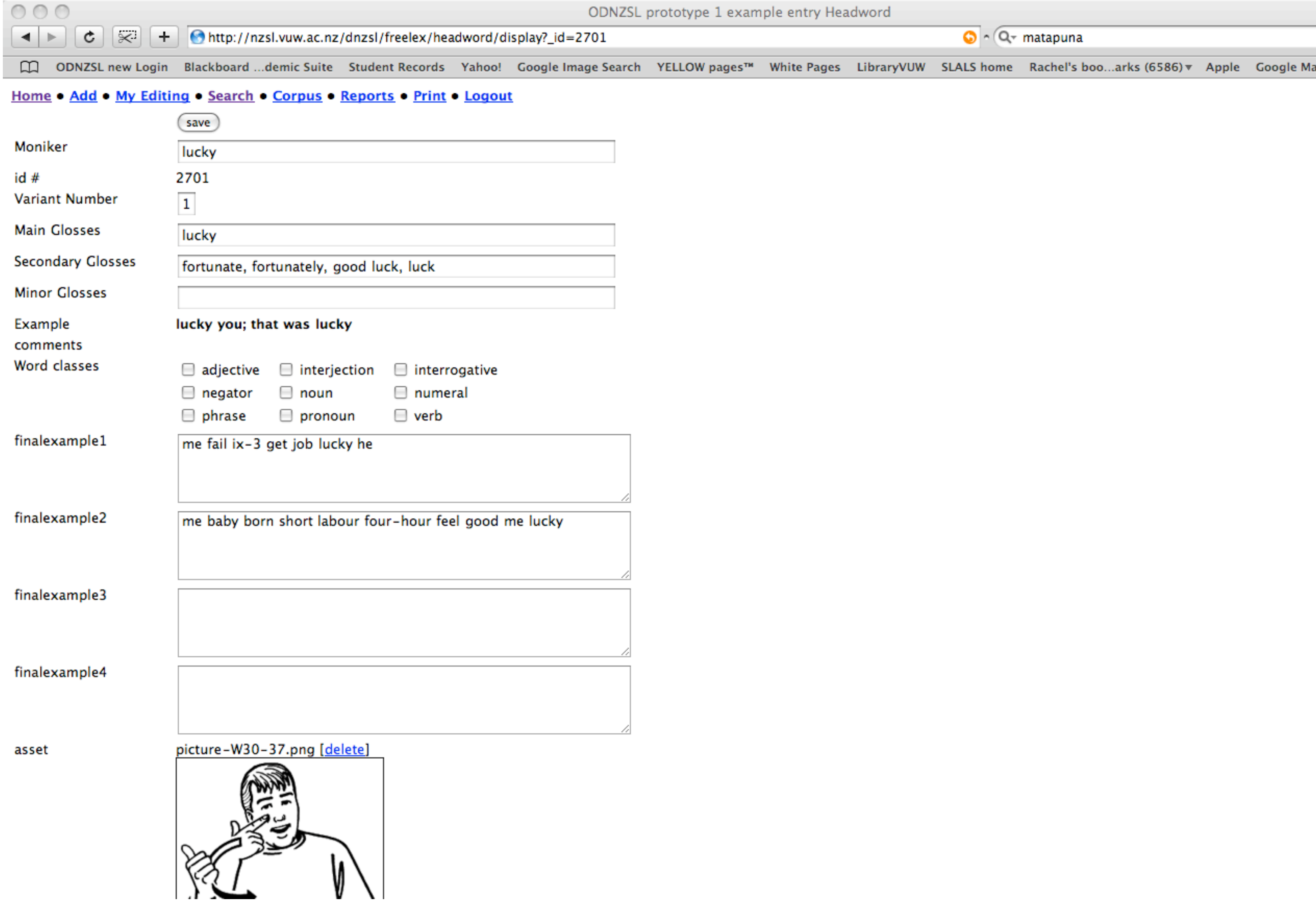
Sociolinguistic Variation archive 2005-2007

- 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

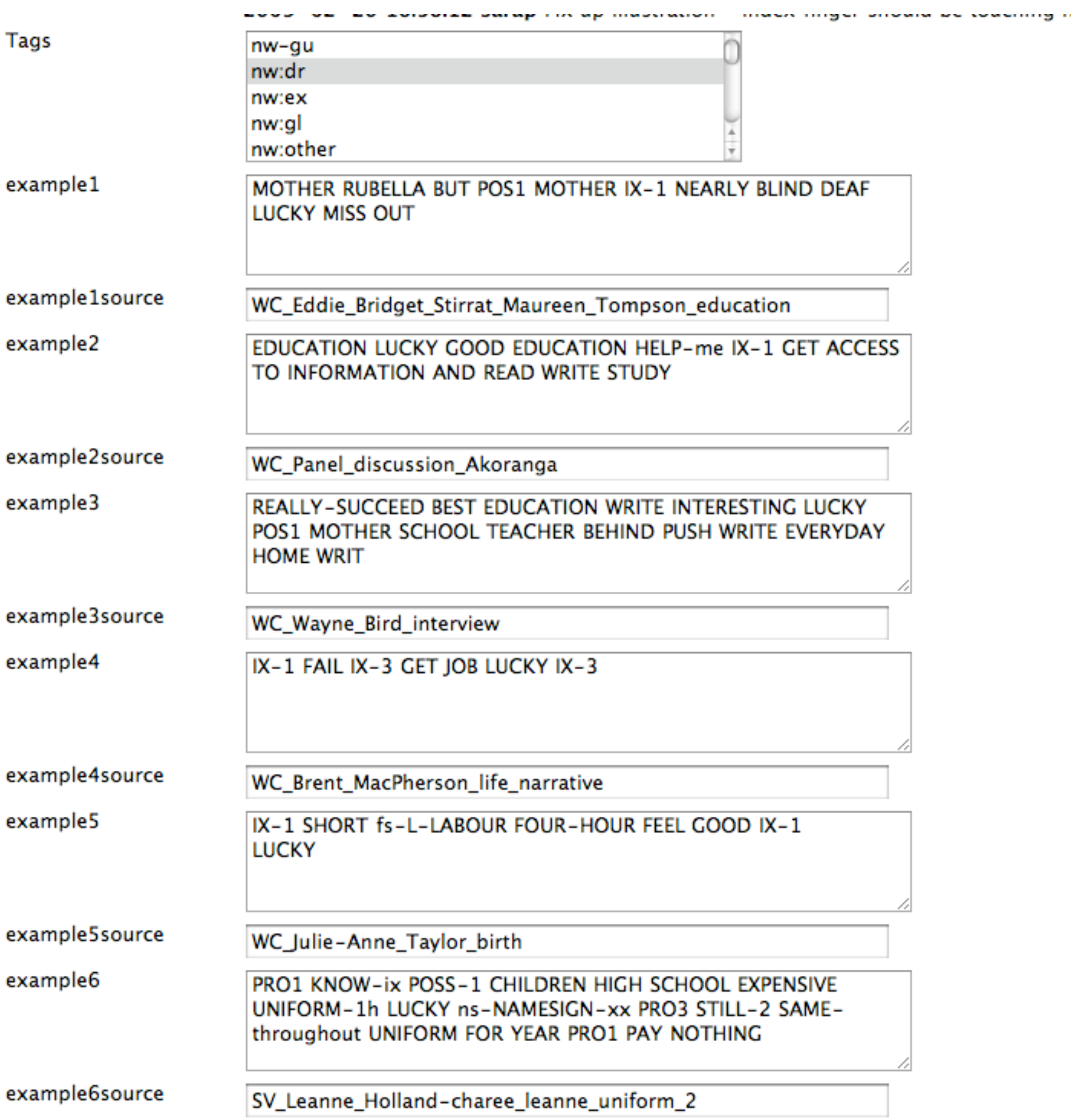
Extending the corpus for use in Online Dictionary

- 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

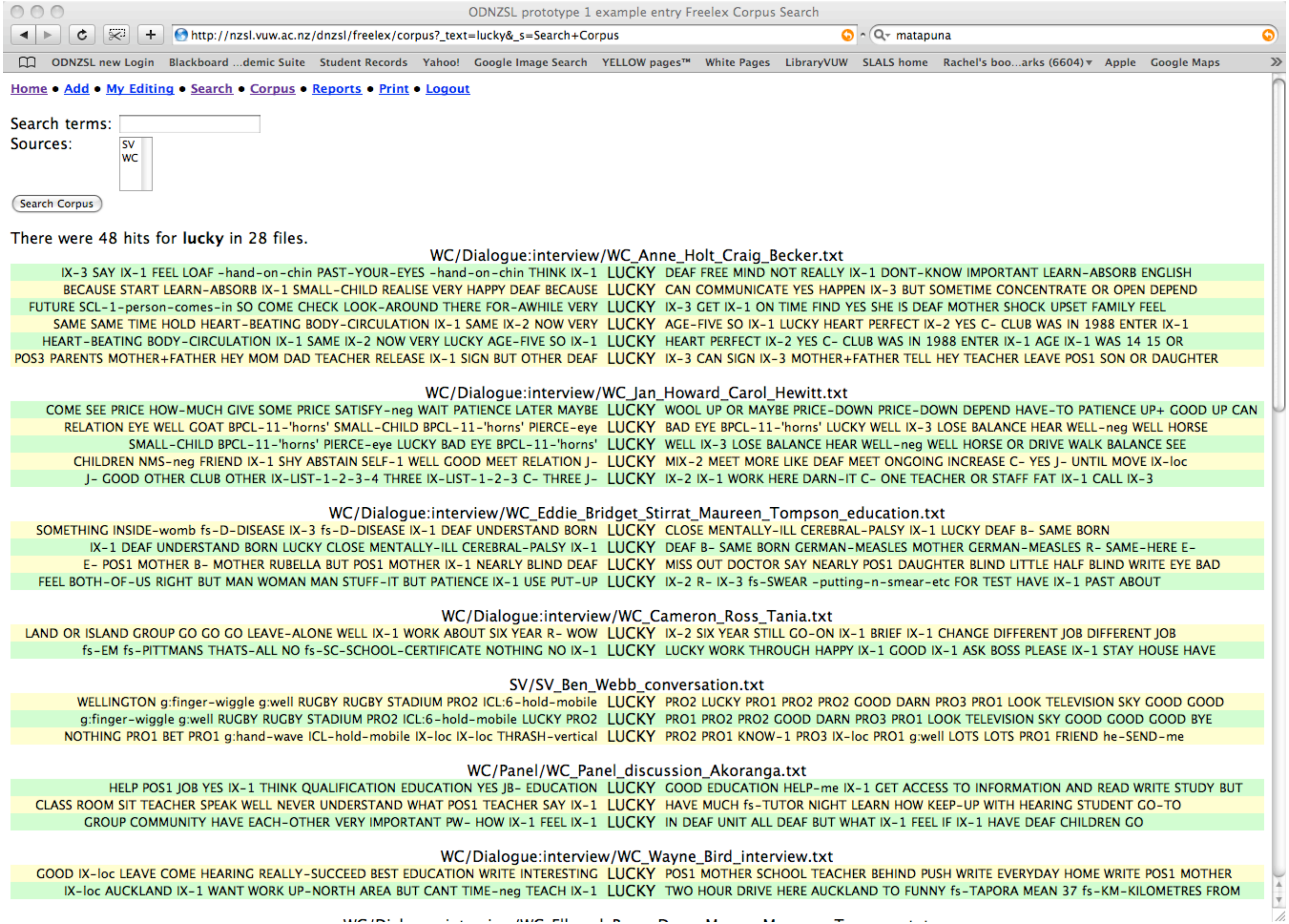

ODNZSL prototype 1 example entry Headword
http://nzsl.vuw.ac.nz/dnzsl/freelex/headword/display?_id=2701
matapuna
ODNZSL new Login
Blackboard ...demic Suite
Student Records
Yahoo!
Google Image Search
YELLOW pages™
White Pages
LibraryVUW
SLALS home
Rachel's boo...arks (6586)
Apple
Google Ma
Home • Add • My Editing • Search • Corpus • Reports • Print • Logout
save

Moniker	lucky
id #	2701
Variant Number	1
Main Glosses	lucky
Secondary Glosses	fortunate, fortunately, good luck, luck
Minor Glosses	
Example	**lucky you; that was lucky**
comments	
Word classes	☐ adjective ☐ interjection ☐ interrogative ☐ negator ☐ noun ☐ numeral ☐ phrase ☐ pronoun ☐ verb
finalexample1	me fail ix-3 get job lucky he
finalexample2	me baby born short labour four-hour feel good me lucky
finalexample3	
finalexample4	
asset	picture-W30-37.png [delete]



2009-02-26 16:36:12 sarap Fix up illustration - index finger should be touching

Tags	nw-gu nw:dr nw:ex nw:gl nw:other
example1	MOTHER RUBELLA BUT POS1 MOTHER IX-1 NEARLY BLIND DEAF LUCKY MISS OUT
example1source	WC_Eddie_Bridget_Stirrat_Maureen_Tompson_education
example2	EDUCATION LUCKY GOOD EDUCATION HELP-me IX-1 GET ACCESS TO INFORMATION AND READ WRITE STUDY
example2source	WC_Panel_discussion_Akoranga
example3	REALLY-SUCCEED BEST EDUCATION WRITE INTERESTING LUCKY POS1 MOTHER SCHOOL TEACHER BEHIND PUSH WRITE EVERYDAY HOME WRIT
example3source	WC_Wayne_Bird_interview
example4	IX-1 FAIL IX-3 GET JOB LUCKY IX-3
example4source	WC_Brent_MacPherson_life_narrative
example5	IX-1 SHORT fs-L-LABOUR FOUR-HOUR FEEL GOOD IX-1 LUCKY
example5source	WC_Julie-Anne_Taylor_birth
example6	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
example6source	SV_Leanne_Holland-charee_leanne_uniform_2



http://nzsl.vuw.ac.nz/dnzsl/freelex/corpus?_text=lucky&_s=Search+Corpus

Home • Add • My Editing • Search • Corpus • Reports • Print • Logout

Search terms:

Sources: SV WC

There were 48 hits for **lucky** in 28 files.

WC/Dialogue:interview/WC_Anne_Holt_Craig_Becker.txt

IX-3 SAY IX-1 FEEL LOAF -hand-on-chin PAST-YOUR-EYES -hand-on-chin THINK IX-1 LUCKY DEAF FREE MIND NOT REALLY IX-1 DONT-KNOW IMPORTANT LEARN-ABSORB ENGLISH

BECAUSE START LEARN-ABSORB IX-1 SMALL-CHILD REALISE VERY HAPPY DEAF BECAUSE LUCKY CAN COMMUNICATE YES HAPPEN IX-3 BUT SOMETIME CONCENTRATE OR OPEN DEPEND

FUTURE SCL-1-person-comes-in SO COME CHECK LOOK-AROUND THERE FOR-AWHILE VERY LUCKY IX-3 GET IX-1 ON TIME FIND YES SHE IS DEAF MOTHER SHOCK UPSET FAMILY FEEL

SAME SAME TIME HOLD HEART-BEATING BODY-CIRCULATION IX-1 SAME IX-2 NOW VERY LUCKY AGE-FIVE SO IX-1 LUCKY HEART PERFECT IX-2 YES C- CLUB WAS IN 1988 ENTER IX-1

HEART-BEATING BODY-CIRCULATION IX-1 SAME IX-2 NOW VERY LUCKY AGE-FIVE SO IX-1 LUCKY HEART PERFECT IX-2 YES C- CLUB WAS IN 1988 ENTER IX-1 AGE IX-1 WAS 14 15 OR

POS3 PARENTS MOTHER+FATHER HEY MOM DAD TEACHER RELEASE IX-1 SIGN BUT OTHER DEAF LUCKY IX-3 CAN SIGN IX-3 MOTHER+FATHER TELL HEY TEACHER LEAVE POS1 SON OR DAUGHTER

WC/Dialogue:interview/WC_Jan_Howard_Carol_Hewitt.txt

COME SEE PRICE HOW-MUCH GIVE SOME PRICE SATISFY-neg WAIT PATIENCE LATER MAYBE LUCKY WOOL UP OR MAYBE PRICE-DOWN PRICE-DOWN DEPEND HAVE-TO PATIENCE UP+ GOOD UP CAN

RELATION EYE WELL GOAT BPCL-11-'horns' SMALL-CHILD BPCL-11-'horns' PIERCE-eye LUCKY BAD EYE BPCL-11-'horns' LUCKY WELL IX-3 LOSE BALANCE HEAR WELL-neg WELL HORSE

SMALL-CHILD BPCL-11-'horns' PIERCE-eye LUCKY BAD EYE BPCL-11-'horns' LUCKY WELL IX-3 LOSE BALANCE HEAR WELL-neg WELL HORSE OR DRIVE WALK BALANCE SEE

CHILDREN NMS-neg FRIEND IX-1 SHY ABSTAIN SELF-1 WELL GOOD MEET RELATION J- LUCKY MIX-2 MEET MORE LIKE DEAF MEET ONGOING INCREASE C- YES J- UNTIL MOVE IX-loc

J- GOOD OTHER CLUB OTHER IX-LIST-1-2-3-4 THREE IX-LIST-1-2-3 C- THREE J- LUCKY IX-2 IX-1 WORK HERE DARN-IT C- ONE TEACHER OR STAFF FAT IX-1 CALL IX-3

WC/Dialogue:interview/WC_Eddie_Bridget_Stirrat_Maureen_Tompson_education.txt

SOMETHING INSIDE-womb fs-D-DISEASE IX-3 fs-D-DISEASE IX-1 DEAF UNDERSTAND BORN LUCKY CLOSE MENTALLY-ILL CEREBRAL-PALSY IX-1 LUCKY DEAF B- SAME BORN

IX-1 DEAF UNDERSTAND BORN LUCKY CLOSE MENTALLY-ILL CEREBRAL-PALSY IX-1 LUCKY DEAF B- SAME BORN GERMAN-MEASLES MOTHER GERMAN-MEASLES R- SAME-HERE E-

E- POS1 MOTHER B- MOTHER RUBELLA BUT POS1 MOTHER IX-1 NEARLY BLIND DEAF LUCKY MISS OUT DOCTOR SAY NEARLY POS1 DAUGHTER BLIND LITTLE HALF BLIND WRITE EYE BAD

FEEL BOTH-OF-US RIGHT BUT MAN WOMAN MAN STUFF-IT BUT PATIENCE IX-1 USE PUT-UP LUCKY IX-2 R- IX-3 fs-SWEAR -putting-n-smear-etc FOR TEST HAVE IX-1 PAST ABOUT

WC/Dialogue:interview/WC_Cameron_Ross_Tania.txt

LAND OR ISLAND GROUP GO GO GO LEAVE-ALONE WELL IX-1 WORK ABOUT SIX YEAR R- WOW LUCKY IX-2 SIX YEAR STILL GO-ON IX-1 BRIEF IX-1 CHANGE DIFFERENT JOB DIFFERENT JOB

fs-EM fs-PITTMANS THATS-ALL NO fs-SC-SCHOOL-CERTIFICATE NOTHING NO IX-1 LUCKY LUCKY WORK THROUGH HAPPY IX-1 GOOD IX-1 ASK BOSS PLEASE IX-1 STAY HOUSE HAVE

SV/SV_Ben_Webb_conversation.txt

WELLINGTON g:finger-wiggle g:well RUGBY RUGBY STADIUM PRO2 ICL:6-hold-mobile LUCKY PRO2 LUCKY PRO1 PRO2 PRO2 GOOD DARN PRO3 PRO1 LOOK TELEVISION SKY GOOD GOOD

g:finger-wiggle g:well RUGBY RUGBY STADIUM PRO2 ICL:6-hold-mobile LUCKY PRO2 LUCKY PRO1 PRO2 PRO2 GOOD DARN PRO3 PRO1 LOOK TELEVISION SKY GOOD GOOD GOOD BYE

NOTHING PRO1 BET PRO1 g:hand-wave ICL-hold-mobile IX-loc IX-loc THRASH-vertical LUCKY PRO2 PRO1 KNOW-1 PRO3 IX-loc PRO1 g:well LOTS LOTS PRO1 FRIEND he-SEND-me

WC/Panel/WC_Panel_discussion_Akoranga.txt

HELP POS1 JOB YES IX-1 THINK QUALIFICATION EDUCATION YES JB- EDUCATION LUCKY GOOD EDUCATION HELP-me IX-1 GET ACCESS TO INFORMATION AND READ WRITE STUDY BUT

CLASS ROOM SIT TEACHER SPEAK WELL NEVER UNDERSTAND WHAT POS1 TEACHER SAY IX-1 LUCKY HAVE MUCH fs-TUTOR NIGHT LEARN HOW KEEP-UP WITH HEARING STUDENT GO-TO

GROUP COMMUNITY HAVE EACH-OTHER VERY IMPORTANT PW- HOW IX-1 FEEL IX-1 LUCKY IN DEAF UNIT ALL DEAF BUT WHAT IX-1 FEEL IF IX-1 HAVE DEAF CHILDREN GO

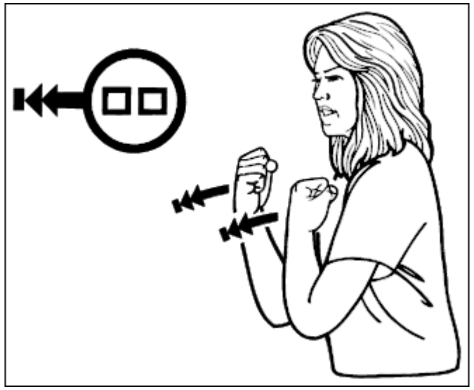
WC/Dialogue:interview/WC_Wayne_Bird_interview.txt

GOOD IX-loc LEAVE COME HEARING REALLY-SUCCEED BEST EDUCATION WRITE INTERESTING LUCKY POS1 MOTHER SCHOOL TEACHER BEHIND PUSH WRITE EVERYDAY HOME WRITE POS1 MOTHER

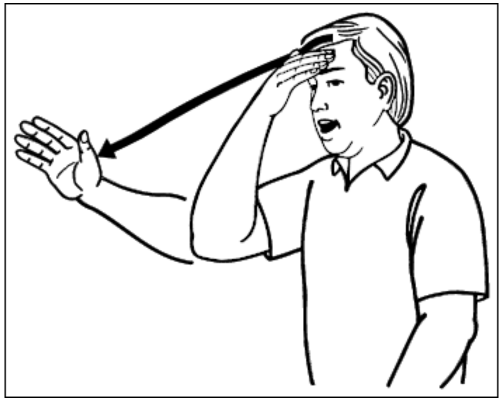
IX-loc AUCKLAND IX-1 WANT WORK UP-NORTH AREA BUT CANT TIME-neg TEACH IX-1 LUCKY TWO HOUR DRIVE HERE AUCKLAND TO FUNNY fs-TAPORA MEAN 37 fs-KM-KILOMETRES FROM

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*

Drawbacks of corpus based examples

- 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

- Atkins, B.T. S. & M. Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford University Press
- Johnston, T. & A. Schembri. 2005. The use of ELAN annotation software in the Auslan Archive/Coprus project. Presentation at the Ethnographic Ereserach Annotaion Conference. Univ of Melbourne.
- McKee, D. & G. Kennedy (1999). A list of 1,000 frequently-used signs in New Zealand Sign Language. In Kennedy, G. (ed.) Topics in New Zealand Sign Language Studies, Deaf Studies Research Un it, Occasional Publication No. 1, Victoria University of Wellington. (17-25).
- McKee, David & Graeme Kennedy 2006. The Distribution of Signs in New Zealand Sign Language. *Sign Language Studies* 6 (4). 372-390
- Morford, J.P & J. MacFarlane (1998). Frequency Characteristics of American Sign Language. *Sign Language Studies*, 3 (2):213-225
- Sinclair, J. 1991 Corpus, Concordance, and Collocation. OUP

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