Automating the Analysis of Large Bilingual Corpora

Diana Carter, Kevin Donnelly
& Mirjam Broersma
diana.carter@ubc.ca
AACL January 20, 2013

Outline

- Motivation
- Siarad Corpus
- Autoglosser
- File preparation
- Clause analysis
- Preparation for data analysis
- Conclusions

Introduction

- Triggering: cognates facilitate CS (Clyne 1967, 2003)
- In the mental lexicon, words are organized in language subsets; activation of one word activates subset (Paradis 1987, 2004)
- Trigger words are part of the subsets of both languages (Broersma & De Bot 2006)
- Selection of trigger word:
  - → causes activation of the other language subset
  - → increases chance of selection of a word in the other language
- Occurs in bilingual mode; not consciously planned

Research Questions

1. What characteristics of cognates affect the extent to which they can facilitate codeswitching?

2. What characteristics of non-cognate words affect their likelihood to undergo cognate-induced codeswitching?

3. How does codeswitching density affect cognate-induced codeswitching?

Siarad Corpus

- 40 hours of spontaneous Welsh-English bilingual speech
- Collected over two years (‘05-’07) in North Wales
- 151 speakers
- 447 507 words
- Transcribed in CLAN using CHAT with audio links
- Publicly available and searchable at http://bangortalk.org.uk

Example of Triggered CS

***ALN:** ond dw i ddim **actually** isio mynd i wrando ar y stuff .

%auto: but.CONJ be.v.1S.PRES I.PRON.1S not.AD+SM actual.ADJ+ADV want.N.M.SG go.V.INF to.PREP listen.V.INF+SM on.PREP the.DET.DEF stuff.N.SG

'but I don't actually want to go and listen to the stuff'

Requirements

- Each file analyzed at the clause-level
  - Presence of a trigger word?
  - If a clause has a trigger word, is there a codeswitch within the clause? (internal) Or in the next clause? (external)
  - What type of word is the trigger word and the switched word?

- Previous studies used manual analysis

Bangor Autoglosser (ESRC Centre)

- One-pass glossing of multilingual conversations
- Uses database for text storage and dictionary lookup
- Uses constraint grammar for disambiguation
- 98% accuracy for Welsh and English
- 1000 words a minute
- GPL license http://bangortalk.org.uk

Donnelly & Deuchar 2011

File Preparation

1. File selection:
   - Selected 52 conversations (out of 69 total)
   - Stored in database as tables as part of the autoglosser process

1. Remove Interactional Markers:
   - Eg. ah, er, mm, mmhm, oh, uh-huh

File Preparation

3. Split into clauses:
   - No parser available for Welsh
   - Used an ad hoc method (95%+ accuracy)
   - Autoglosser produced a database table with part-of-speech tags assigned to each word
   - In the db table, added a marker at main (finite) verbs in each utterance, moved it where appropriate (conjunctions, relatives etc)
   - Split the utterance into clauses at the marker

Table with vertical text

location	surface	auto	langid
1	ond	but.CONJ	cym
2	dw	be.V.1S.PRES	cym
3	i	I.PRON.1S	cym
4	ddim	not.ADV+SM	cym
5	actually	actual.ADJ+ADV	eng
6	isio	want.N.M.SG	cym
7	mynd	go.V.INFIN	cym
8	i	to.PREP	cym
9	wrando	listen.V.INFIN+SM	cym
10	ar	on.PREP	cym
11	y	the.DET.DEF	cym
12	stuff	stuff.N.SG	cym&eng
13	.	*NULL*	999

File Preparation

4. Mark cognate words:
   - Nouns, adjectives, adverbs, exclamations, names, verbs based on English

5. Mark speaker-turns:
   - Aggregated clauses into blocks by each speaker
   - Ignored speaker-turns consisting solely of minimal-content items (ie iawn/ok, na/no, timod/you know, yeah, ydy/isn't it)

Clause Analysis

1. Create a new table based on speaker-turns

2. Generate additional data about the clause:
   - Enrich existing info by drawing on/combining it to specify further variables

1. Use the additional data to locate and characterize codeswitches:
   - Only non-cognates are used to determine CSs
   - External / Internal

Clause Analysis

4. Combine this with whether or not a cognate occurs in the clause:
   - ST: codeswitch, and cognate (trigger) is present
   - NST: no codeswitch, cognate is present
   - SNT: codeswitch, no cognate
   - NSNT: no codeswitch, no cognate

4. With some clauses, none of these categories may apply

Preparation for Data Analysis

1. Create a new table including the codeswitching data:
   - Generate additional info where necessary (eg. length of cognate)
   - If more than one cognate occurs in one clause, create multiple records
1. Include frequency data:
   - Clauses, words, triggers, codeswitches
   - For each speaker
   - For the file as a whole

Preparation for Data Analysis

3. Create a new file ready for input into R for stats analysis:
   - Combine all the data for each conversation into one file
   - Convert to CSV

Conclusions

- Pipeline
  - Allows incremental development
  - Easy to isolate and check output at each stage (via intermediate tables)
  - Revisions can be made at any stage with little impact on other stages
  - Various pipeline stages can be automated from a shell script

Conclusions

- Using a scripting language like PHP
  - Shallower learning curve
  - Immediate feedback (no compilation required)
  - Transferable skills
  - No need to learn specific interface for a standalone app (eg CLAN)

Diolch
Thank you

References

Broersma, M., & De Bot, K. (2006). Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9, 1-13.

Clyne, M. (2003). *Dynamics of language contact: English and immigrant languages.* Cambridge: Cambridge University Press.

Donnelly, K., & Deuchar, M. (2011). Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text. In *Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop*, Riga, Latvia. NEALT Proceedings Series, Tartu.

References

Deuchar, M., Davies, P., Herring, J., Parafita Couto, M.C., and D. Carter. (forthcoming). Building Bilingual Corpora. In E. Thomas and I. Mennen (eds.) *Unraveling Bilingualism: A Cross-Disciplinary Perspective*. Bristol: Multilingual Matters.

Deuchar, M., Davies, P., & Donnelly, K. (in prep). Building and using the Siarad Corpus of Spoken Welsh: Bilingual conversations in Welsh and English. John Benjamins Publishers.

Paradis, M. (2004). *A neurolinguistic theory of bilingualism.* Amsterdam: John Benjamins.