Less-than-chance Similarity & Language Differentiation

T. Mark Ellison & Luisa Miceli

Overview

- Introduction & description of research project
- Australian languages as the initial inspiration
- Contact-induced lexical differentiation
- Methodology
- Test case & preliminary results
- Future directions

Introduction

- Differentiation as a result of internal change – we know the historical signature well.
- Contact-induced change – different historical signatures depending on type of situation and intensity of contact.
- Most work on contact-induced change has focused on change that leads to increased similarity.
- Less is known about contact-induced change that leads to differentiation.
- This is the focus of our project.

Broad description of project

- As just mentioned, the focus of this project is contact-induced differentiation and, in particular, its historical signature.
- Our hypothesis is that this type of differentiation leads to less-than-chance similarity.
  - Stage 1: Development of a methodology to measure linguistic similarity (lexicon)
  - Stage 2: Testing on reported cases of contact-induced lexical differentiation
  - Stage 3: Diagnosis of prehistoric instances

Initial inspiration for the project

- Australian languages – in particular, the mismatch between degree of structural and lexical similarity:
    - much structural similarity
    - little lexical similarity
- Our hypothesis is that at least in those cases where the mismatch is most extreme (e.g. some Northern Australian languages) there may have been contact-induced lexical differentiation.

‘Traditional’ explanations of the mismatch

- Contact has led to high degrees of structural similarity.
  - But why not more lexical borrowing?
- Higher than expected rates of lexical replacement have led to comparatively less lexical similarity in comparison to structural similarity.
  - Due to practices such as death-taboo – but not evident in the few historical wordlists available (Alpher & Nash 1999).
  - And, in any case, this type of motivation for replacement is language internal.

Explanation we are investigating

- Both the high degree of structural similarity and the low degree of lexical similarity are due to contact.
- Contact-induced lexical differentiation:
  - For a given meaning, when there are several forms available, preference is given to the synonym less similar in form to that in the other language(s) in the linguistic repertoire – avoidance of cognates & lexical look-alikes.
  - Avoidance of borrowing as a means for lexical replacement.
    - This second possibility was also discussed in Harvey (2006)

Does contact-induced lexical differentiation actually occur?

- It has been reported in a number of multilingual speech communities in different parts of the world.
- Contact-induced differentiation is not limited to the lexicon, but predominantly affects phonology and lexicon (Thomason 2007).

Laycock (1982): Uisai

- "… Melanesian exploitation of diversity … evidence that additional difference is created."
- "In [the Uisai dialect of Buin] … we find all the gender agreements reversed … all the masculines are feminine and all the feminines are masculine. There is no accepted mechanism for linguistic change which can cause a flip-flop of this kind and magnitude." (p.36)

Trudgill (1986): ‘r-ful’ dialects in England

- ‘r-ful’ dialects bordering onto ‘r-less’ dialects in England, insert post-vocalic ‘r’ in a number of words that etymologically had no ‘r’:
  - e.g. walk, calf, straw, daughter etc.

Beswick (2007): 19th Century Galician

- "…popular words shared with Castilian were either rejected in favour of Galician synonyms or phonetically or morphologically altered through a process of *hyperpurism*." (p.116)

Wright (1998): present day Catalan & Galician

- “where Catalan, or Galician, has two words that are for practical purposes synonymous, one which is like Castilian, one which is not, the dictionary and standardizers … have tended to prefer the one which is not like Castilian.”

Fabra (1924-25): Catalan

- *“Hi hagué una època … en tota coincidència entre l’espanyol i el català, es veia un castellanisme, i bastava que un mot s’assemblès massa a l’espanyol correspondent perquè se li cerquès … un substitut.”* (p.16)

  The was a time when … in every agreement between Spanish and Catalan a castilianism was seen, and a word only had to look too similar to the corresponding Spanish one in order for … substitutions for it to be sought. (translation, Carrasquer Vidal 1998)

- Carrasquer Vidal points out that in the above passage itself, there are two examples of differentiation!

Fabra (1924-25): Catalan

- *“Hi hagué una època … en tota coincidència entre l’espanyol i el català, es veia un castellanisme, i bastava que un mot s’assemblès massa a l’espanyol correspondent perquè se li cerquès … un substitut.”* (p.16)
  - *mots* instead of *paraules*
  - *cerquès* instead of *busquis*

Carrasquer Vidal (1998): spoken Catalan

- Admits that many Castilianisms still exist in spoken Catalan.
- But that the number has been drastically reduced.

Motivations for contact-induced differentiation

- Obvious from discussed examples, that contact-induced differentiation often falls into the category of ‘deliberate’ change.
- Usually occurs when there is either:
  - a desire or need to increase the difference between one’s own speech and someone else’s.
  - a desire to keep outsiders at a linguistic distance. (Thomason 2007)

A possible motivation for contact-induced *lexical* differentiation specifically

- In a sociolinguistic setting where more than one language is used on a daily basis:
  - does lexical differentiation ease the cognitive burden of the individual speaker?

Relevant psycholinguistic findings

- Interlingual homophones are harder to process than words that belong exclusively to one language. (Grojean 1988)
- Schulpen, Dijkstra, Schriefers & Hasper (2003), same effect as Grosjean - word identification and language membership decisions by Dutch-English bilinguals were delayed for interlingual homophones.

So, perhaps, as a response to the heavy cognitive load …

Unrelated languages

structure converges

lexicon maintained distinct and differentiated

(avoidence of borrowing & lexical look-alikes)

Related languages

structural similarity maintained (& change affects all languages in the repertoire)

lexicon undergoes differentiation

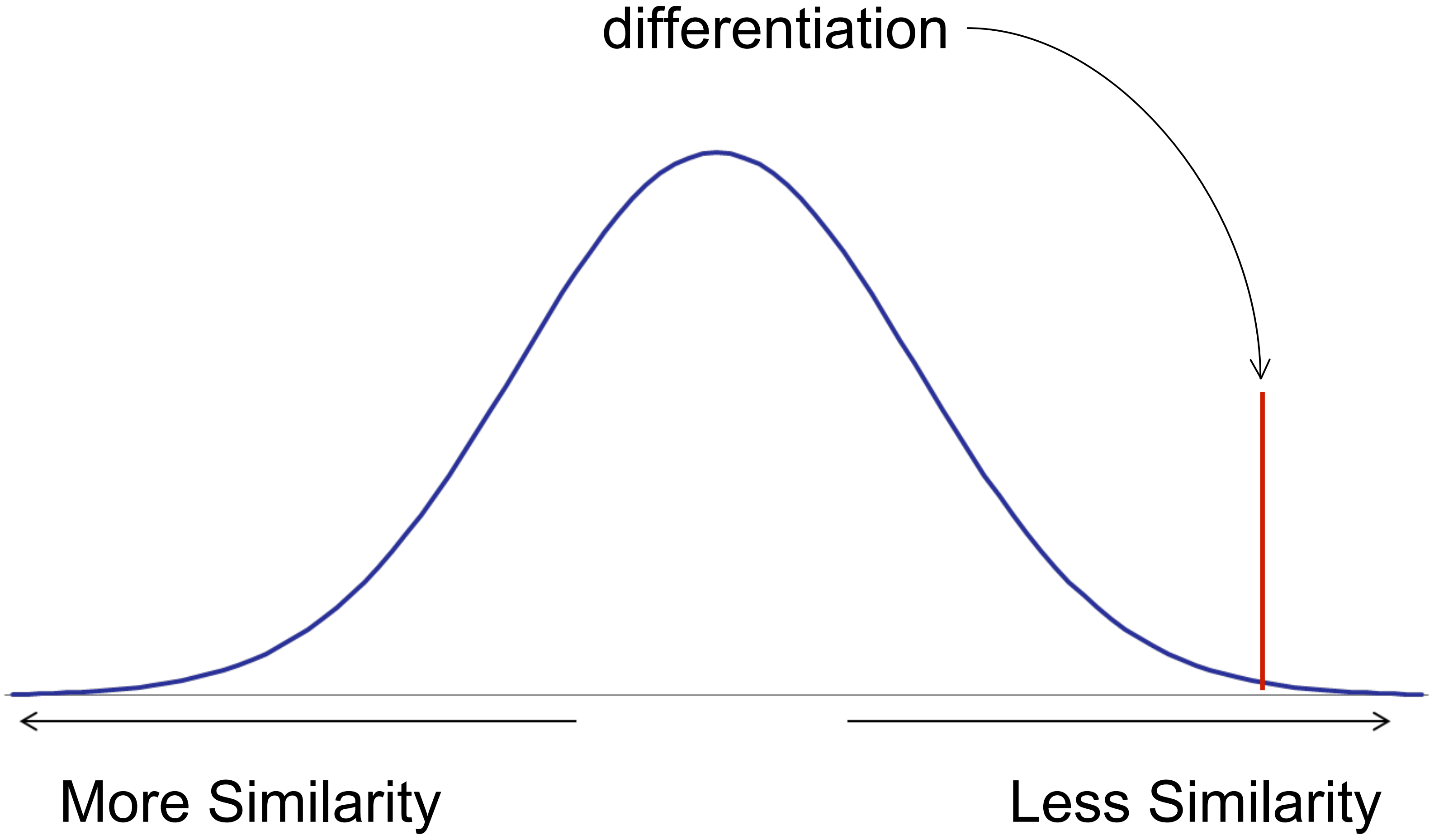
The historical signature of contact-induced lexical differentiation

- As mentioned earlier, our hypothesis is that contact-induced lexical differentiation gives rise to less-than-chance similarity in the lexicon.
- Mark will now describe the method that we have been developing to measure linguistic similarity.
- And demonstrate its application using Catalan/Castillian data.

Identifying Past Differentiation

- our long-term goal is a method to identify past differentiation
- given synchronic data
  - eg dictionaries, wordnet, corpora
- by comparing actual similarity to what we would expect by chance
- will illustrate what we have so far with Castillian and Catalan

Unlikely Dissimilarity



Catalan and Castillian Data

- wordnets for Catalan, Castillian*
- *wordnet* – a lexical database with:
  - synsets – senses/meanings
    - same as English wordnet synsets
  - variants – forms expressing these senses
  - relations – hypernym, meronym, etc.
- we use synsets and their variants

*http://www.lsi.upc.edu/~nlp/web/index.php?option=com_content&task=view&id=31&Itemid=57

SynSets

Catalan

ʑesta
feta
fita
konsekusio

Castillian

aθaɲa
konsekuθion
logro
proeθa
xesta

Segment Similarity

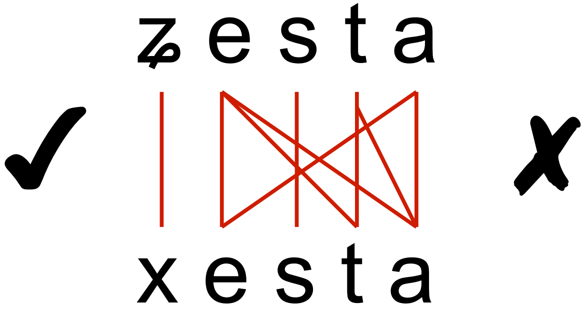
- union of the segment inventories of the two languages
- confusion probability (CP) over pairs of segments
  - based on overlapping features
  - adjusted for segment frequency

a~a 0.066, **m~n** 0.029, **i~i** 0.053,

s~θ 0.027, **s~Ø** 0.016, ...

Alignment Similarity

- an alignment maps segments of one word to segments of another such that:
  - mappings do not cross
  - no segment has more than one mapping



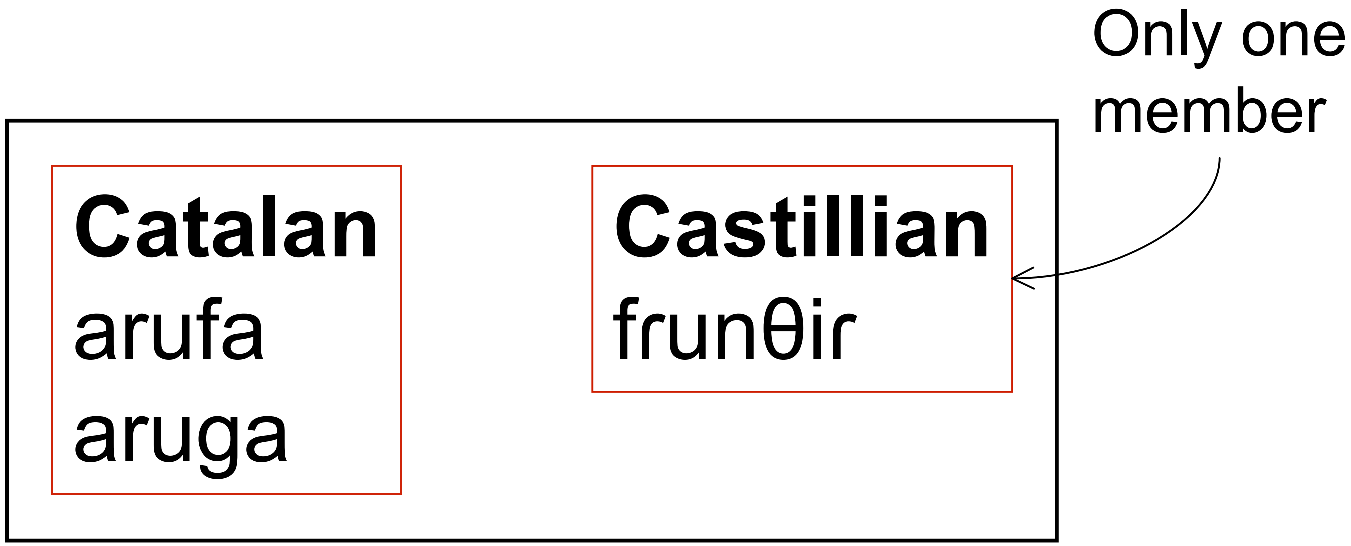
- product CPs of aligned pairs, or zero

Word-Word Similarity

- sum the alignment similarities for every possible alignment of the two words
- there are very many alignments
  - but can adapt algorithms for computing Levenshtein distances to make feasible
- similarities are scaled by word lengths
  - so long words can be as similar as short

Singleton Synsets

- synset *size* counts Castillian words
- a *singleton* synset is one with size 1



Only one member

Catalan
arufa
aruga

Castillian
frunθir

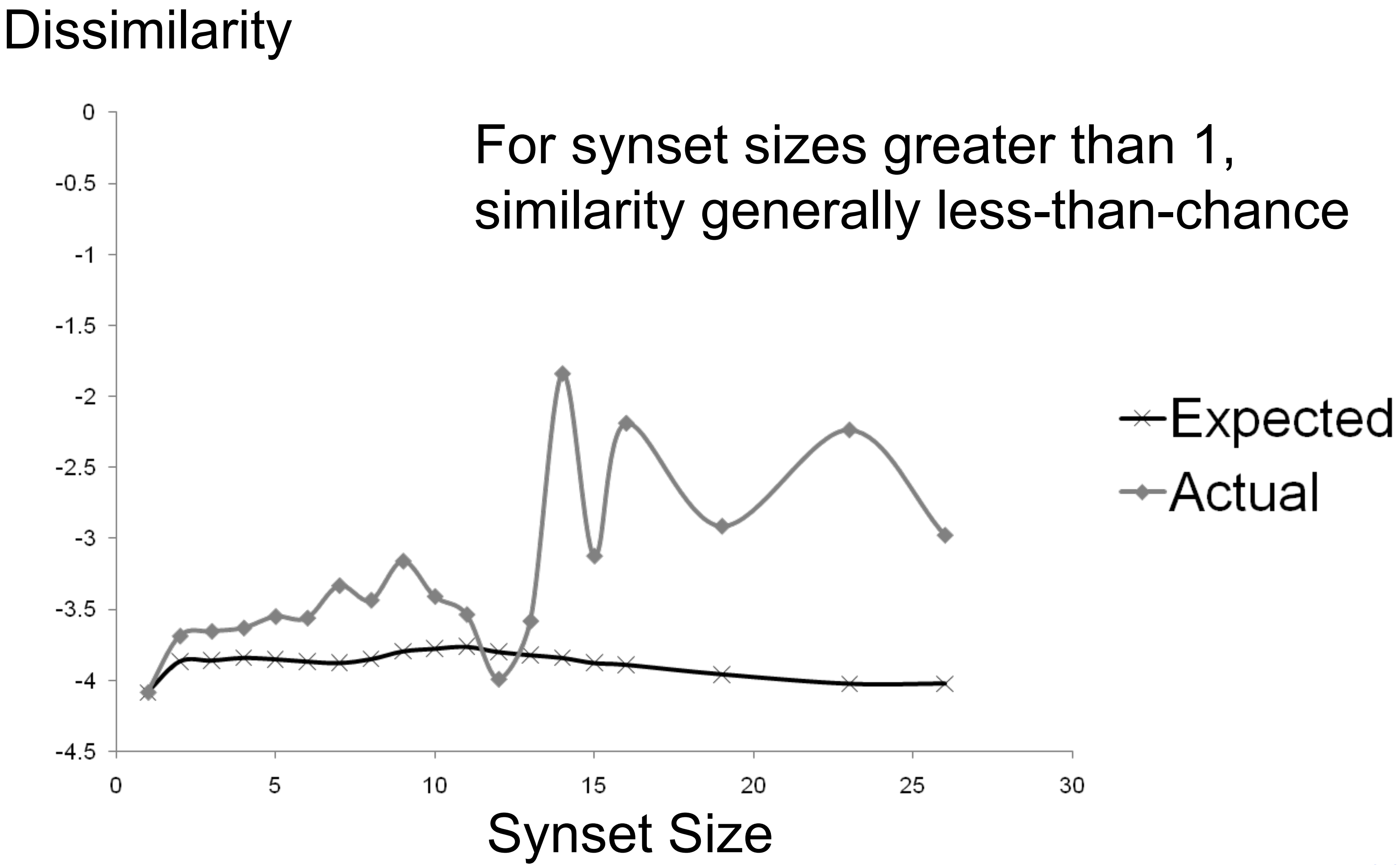
Non-Singleton Synsets

- have multiple Castillian word forms
- for each word
    - measure its similarity to the *most similar* corresponding word in the other language
        - is likely to match words with a cognate
- aggregate similarities with those in other synsets of the same size

Expected Similarity of Non-Singleton Synsets

- computed from singleton synset similarities
- pick random *n* singleton synsets
- treat variants from these as if from one big synset
- compute the similarities
- repeat, to compute expected average similarity for synsets of size *n*

Results



For synset sizes greater than 1, similarity generally less-than-chance

Conclusion

- strong anecdotal evidence that differentiation does occur
- in Catalan vs Castillian
  - seems to be a choice between synonyms
  - reflected statistically with less-than-chance similarity
- the method can find statistical evidence for past differentiation

Future Work

- look at a control case
- richer similarity models, eg. HMMs
- explore psycho- and socio-linguistic factors triggering differentiation
- more detailed analysis of case studies
- look at new data
    - do you have some?

References

- Atserias, J., Climent, S., Farreres, J., Rigau, G & H. Rodríguez. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets Proceedings of Conference on Recent Advances on NLP. RANLP 97. Tzigov Chark, Bulgaria, 1997.
- Alpher, Barry & David Nash 1999. Lexical replacement and cognate equilibrium in Australia. *Australian Journal of Linguistics,* 19, pp. 5-56.
- Beswick, Jaine E. 2007. *Regional Nationalism in Spain: language use and ethnic identity in Galicia*. Clevedon: Multilingual Matters.
- Carrasquer Vidal, Miguel. 1998. Unititled post in the thread 'Cladistic language concepts', HISTLING list.
- Fabra, Pompeu. 1924-25. L'obra de depuraciódel català (discurs llegit en la sessió inaugural del curs acadèmic de 1924-25). Barcelona: Ateneu Barcélonès.
- Grosjean, F. 1988. Exploring the recognition of processes of guest words in bilingual speech. *Perception & Psycholinguistics*, 28, pp.267-283.
- Harvey, Mark 2006. Lexical change in pre-colonial Australia, paper presented at the Australian Linguistic Society Conference.
- Laycock, Donald C. 1982. Melanesian linguistic diversity: a Melanesian choice? In *Melanesia: beyond diversity*, ed. by R.J. May & H. Nelson, 33-38. Canberra: Australian National University Press.
- Schulpen, Béryl, Dijkstra, Ton, Schriefers, Herbert J. & Mark Hasper 2003. Recognition of interlingual homophones in bilingual auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 29, pp.1155-1178.
- Thomason, Sarah Grey 2007. Language contact and deliberate change. *Journal of language contact*, Thema 1, pp. 41-62
- Trudgill, Peter 1986. *Dialects in contact*. Oxford: Basil Blackwell.
- Wright, Roger 1998. Unpublished post in the thread 'Cladistic language concepts', HISTLING list.

Thanks to …

- We wish to thank the Natural Language Processing Research Group, University of Barcelona, for making Catalan and Spanish Wordnets available to us. http://www.lsi.upc.edu/~nlp/web/index.php?option=com_content&task=view&id=31&Itemid=57