## Class 2, Variation II: MaxEnt OT, lexical selection

**To do for next time**
- Anttila study questions due tomorrow (Friday), to my mailbox in Campbell 3125, by 4 PM. Paper only!
- Anderson ch. 9 study questions due Tuesday
- First assignment (modeling variation) will be posted by tonight; due Friday, Jan. 20.

**Overview:** Last time we saw Stochastic OT for handling free variation. We introduce a rival model, Maximum Entropy OT. Then, we discuss lexical selection.

### 1. Recall our schematic example of free variation in Stochastic OT

| /θɪk/ | *θ<br>ranking value: 101 | IDENT(cont)<br>ranking value: 99 |
|---|---|---|
| wins ~90% of time *a* [θɪk] | * | |
| wins ~10% of time *b* [t̪ɪk] | | * |

The two ranking values are the right distance apart so that, after adding noise to the ranking values, *θ >> IDENT(cont) 90% of the time.

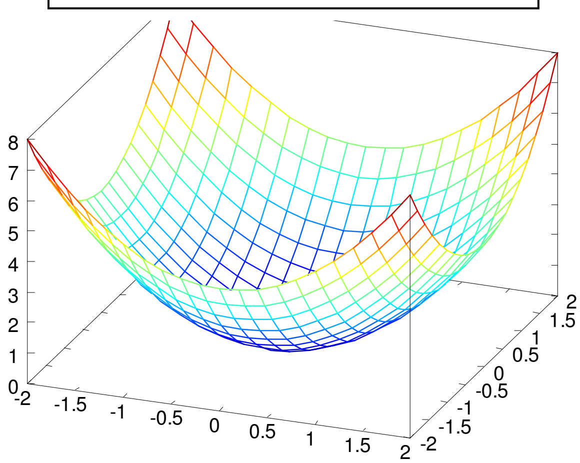
Algorithm for learning these numbers (Gradual Learning Algorithm) demotes and promotes constraints in response to mismatches between current grammar and an adult's utterance.

### 2. Convergence

Suppose for simplicity that our job is to assign numbers (like ranking values) to just two constraints (horizontal axes), to minimize some error rate that we can quantify (vertical axis).

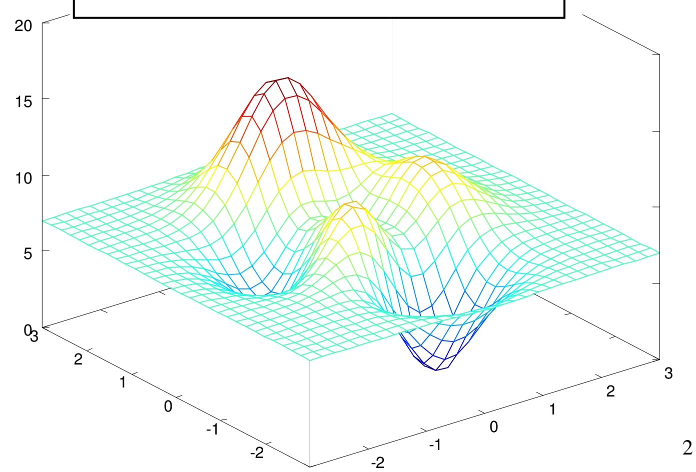
Suppose our learning algorithm goes downhill from wherever we start.

Wherever we start from, we'll end up at the lowest point:




[1]

Here, we risk getting stuck in the shallower pit:




[2]

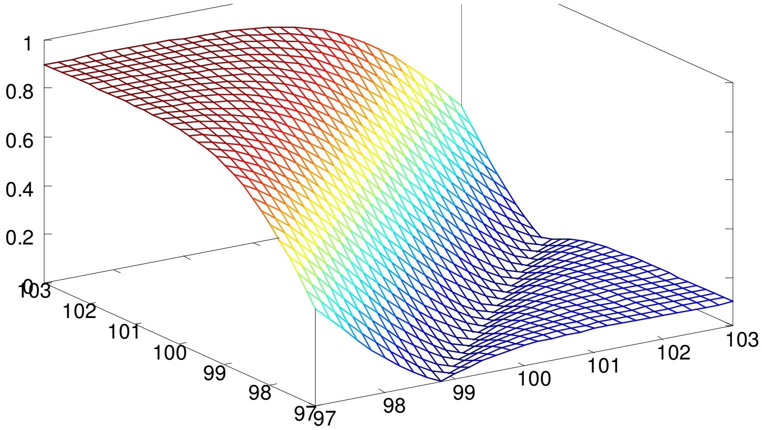
---

[1] Octave commands: [x,y]=meshgrid([-2:.2:2]); Z=x.^2+y.^2; mesh(x,y,Z) . Based on www.mathworks.com/help/techdoc/visualize/f0-18164.html

[2] [X,Y,Z]=peaks(30); mesh(X,Y,Z+7) . Based on www.mathworks.com/help/techdoc/ref/surfc.html

- Ideally, we want a situation as on the left: there's exactly one point from which every direction is uphill. A learning problem like this is called "convex".
- And, we want to know that our learning algorithm always goes downhill, and so is guaranteed to find that optimum.

By the way, here's the plot for our 'thick' example. Error measure is | predicted[t̪]rate – 0.9 |:




[3]

## 3. Gradual Learning Algorithm?

There was no proof of convergence for the standard GLA. Pater 2008 pointed out a type of case that is problematic (the "credit problem").

Suppose 6 constraints: NoCoda, Onset, *VoicedObstruent, *ɛ#, Max-C, Dep-C

○ Determine the constraint ranking for this language and fill in the tableaux

| /da/ | | | | | | |
|---|---|---|---|---|---|---|
| ☞ *a* da | | | | | | |
| *b* a | | | | | | |
| /lob/ | | | | | | |
| *c* lob | | | | | | |
| ☞ *d* lo | | | | | | |
| /tɛf/ | | | | | | |
| ☞ *e* tɛf | | | | | | |
| *f* tɛ | | | | | | |
| /kɛ/ | | | | | | |
| *g* kɛʔ | | | | | | |
| ☞ *h* kɛ | | | | | | |

○ Let's think about what the GLA will do for each of these tableaux if it gets it wrong

⇒ Now let's try it in OTSoft (I have an input file prepared)

(But see www.fon.hum.uva.nl/paul/gla/ for bibliography and discussion: there are variants of GLA that don't have this problem.)

[3] [x,y]=meshgrid([97:0.2:103]); Z=abs(0.9-normcdf(x-y,0,sqrt(2))); mesh(x,y,Z).

## 4. Maximum Entropy OT (Goldwater & Johnson 2003)

Another way to attach numbers to constraints; produces a convex learning situation. Call each constraint's number its "weight".

probability of choosing candidate *x*

$$P(x) = \frac{e^{-\sum_{i=1}^{N} w_i C_i(x)}}{Z}$$

for all *N* constraints, sum of constraint's weight * how many times candidate *x* violates that constraint

sum of these numerators for all the candidates

Example [I cheated and took out DEP-C]

| | ONSET weight: 50 | *VOICEDOBS weight: 19.9 | MAX-C weight: 23.7 | NOCODA weight: 16.4 | *ɛ# weight: 4.5 |
|---|---|---|---|---|---|
| ☞ *a* /da/ → da | | * | | | |
| *b* /da/ → a | * | | * | | |
| *c* /lob/ → lob | | * | | * | |
| ☞ *d* /lob/ → lo | | | * | | |
| ☞ *e* /tɛf/ → tɛf | | | | * | |
| *f* /tɛf/ → tɛ | | | * | | * |
| *g* /kɛ/ → kɛʔ | | | | * | |
| ☞ *h* /kɛ/ → kɛ | | | | | * |

$$P(tEf) = \frac{e^{-(50*0+19.9*0+23.7*0+16.4*1+4.5*0)}}{e^{-(50*0+19.9*0+23.7*0+16.4*1+4.5*0)} + e^{-(50*0+19.9*0+23.7*1+16.4*0+4.5*1)}} = 0.9999925$$

<u>How are the weights chosen</u>?

- So that its predicted probabilities for the correct outputs are as large as possible
- More precisely, maximize the sum of the logs of the predicted probabilities of the *M* pieces of data: $\sum_{j=1}^{M} \ln P(x_j)$ (this is where the "entropy" part of the name comes from)
- You can also include a "smoothing term" to penalize weights far from default value

<u>How are the weights learned</u>?

- OTSoft (and other software) will do it for you, using the Conjugate Gradient Algorithm (see Shewchuk 1994 for tutorial), a fancy version of rolling downhill.

<u>What about free variation</u>?

- Suppose /da/ occurs 10 times, 90% [da], 10% [a].
- If we have weights that produce 99% [da], sum of log probabilities is ln(.99+.99+.99+.99+.99+.99+.99+.99+.99+.01) = -4.696
- But if we have weights that produce 90% [da] (matching the rate in the data), sum of log probabilities is ln(.90+.90+.90+.90+.90+.90+.90+.90+.90+.10) = -3.251, which is bigger.

In your first weekly assignment you'll apply Stochastic OT and MaxEnt OT to the same data set and make some comparisons. Now, let's change topics...

## 5. Lexical selection

Consider English monosyllables beginning sC and ending with a C, sC{l,ɹ,w,j}*V{l,ɹ,[+nas]}CC*#, as listed in CMU pronouncing dicitonary:[4]

| $C_2$ \ $C_1$ | p | b | f | v | m | θ | t | d | s | z | n | l | ɹ | tʃ | dʒ | ʃ | k | g | ŋ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | | | 3 | | 3 | 3 | 39 | 20 | 14 | 12 | 35 | 27 | 21 | 1 | 5 | 2 | 36 | 6 | 9 |
| m | | | 2 | | | 5 | 12 | 3 | 1 | | | 12 | 5 | | 2 | 2 | 13 | 2 | |
| t | 55 | 25 | 26 | 18 | 30 | 2 | 66 | 31 | 11 | 20 | 39 | 44 | 34 | 13 | 9 | 2 | 80 | 7 | 15 |
| n | 11 | 4 | 6 | | | 1 | 4 | 4 | | 4 | | 6 | 8 | 3 | | | 12 | 5 | |
| l | 20 | 4 | 3 | 8 | 9 | 4 | 20 | 10 | 5 | 3 | 7 | | | 1 | 2 | 5 | 8 | 6 | 4 |
| k | 32 | 9 | 16 | 2 | 14 | | 33 | 19 | 5 | 16 | 19 | 28 | 20 | 14 | 4 | 2 | 13 | 8 | |
| w | 24 | 2 | 3 | 3 | 9 | 1 | 15 | 8 | 4 | 5 | 14 | 7 | 5 | 5 | | 4 | 4 | 2 | 3 |

- Certain areas of the chart are underpopulated—discuss.

How underpopulated? Compare to what we expect if each combination depends just on row and column totals:

| $C_2$ \ $C_1$ | p | b | f | v | m | θ | t | d | s | z | n | l | ɹ | tʃ | dʒ | ʃ | k | g | ŋ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 24.1 | 7.5 | 10.0 | 5.3 | 11.1 | 2.7 | 32.1 | 16.5 | 7.0 | 10.2 | 19.4 | 21.1 | 16.3 | 6.3 | 3.7 | 2.9 | 28.4 | 6.1 | 5.3 |
| m | 6.0 | 1.9 | 2.5 | 1.3 | 2.8 | 0.7 | 8.0 | 4.1 | 1.7 | 2.6 | 4.8 | 5.3 | 4.1 | 1.6 | 0.9 | 0.7 | 7.1 | 1.5 | 1.3 |
| t | 53.9 | 16.7 | 22.4 | 11.8 | 24.7 | 6.1 | 71.8 | 36.8 | 15.6 | 22.8 | 43.3 | 47.1 | 36.4 | 14.0 | 8.4 | 6.5 | 63.4 | 13.7 | 11.8 |
| n | 7.0 | 2.2 | 2.9 | 1.5 | 3.2 | 0.8 | 9.3 | 4.8 | 2.0 | 2.9 | 5.6 | 6.1 | 4.7 | 1.8 | 1.1 | 0.8 | 8.2 | 1.8 | 1.5 |
| l | 12.2 | 3.8 | 5.1 | 2.7 | 5.6 | 1.4 | 16.2 | 8.3 | 3.5 | 5.1 | 9.8 | 10.6 | 8.2 | 3.2 | 1.9 | 1.5 | 14.3 | 3.1 | 2.7 |
| k | 26.0 | 8.1 | 10.8 | 5.7 | 11.9 | 2.9 | 34.6 | 17.8 | 7.5 | 11.0 | 20.9 | 22.7 | 17.6 | 6.8 | 4.0 | 3.1 | 30.6 | 6.6 | 5.7 |
| w | 12.1 | 3.7 | 5.0 | 2.6 | 5.5 | 1.4 | 16.1 | 8.2 | 3.5 | 5.1 | 9.7 | 10.5 | 8.2 | 3.1 | 1.9 | 1.4 | 14.2 | 3.1 | 2.6 |

Now take the ratio Observed/Expected—see Frisch, Pierrehumbert, & Broe 2004. I removed cells where Expected < 5, and shaded cells where O/E ≤ 0.5:

| $C_2$ \ $C_1$ | p | b | f | v | m | θ | t | d | s | z | n | l | ɹ | tʃ | dʒ | ʃ | k | g | ŋ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 0.0 | 0.0 | 0.3 | 0.0 | 0.3 | | 1.2 | 1.2 | 2.0 | 1.2 | 1.8 | 1.3 | 1.3 | 0.2 | | | 1.3 | 1.0 | 1.7 |
| m | 0.0 | | | | | | 1.5 | | | | | 2.3 | | | | | 1.8 | | |
| t | 1.0 | 1.5 | 1.2 | 1.5 | 1.2 | 0.3 | 0.9 | 0.8 | 0.7 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 1.1 | 0.3 | 1.3 | 0.5 | 1.3 |
| n | 1.6 | | | | | | 0.4 | | | | 0.0 | 1.0 | | | | | 1.5 | | |
| l | 1.6 | | 0.6 | | 1.6 | | 1.2 | 1.2 | | 0.6 | 0.7 | 0.0 | 0.0 | | | | 0.6 | | |
| k | 1.2 | 1.1 | 1.5 | 0.4 | 1.2 | | 1.0 | 1.1 | 0.7 | 1.5 | 0.9 | 1.2 | 1.1 | 2.1 | | | 0.4 | 1.2 | 0.0 |
| w | 2.0 | | 0.6 | | 1.6 | | 0.9 | 1.0 | | 1.0 | 1.4 | 0.7 | 0.6 | | | | 0.3 | | |

[4] grep ' S [^AEIOU][^AEIOU]*[AEIOU][AEIOU]*[^AEIOU]*[^AEIUO]$' cmudict_0_6d.txt
Excluded row and columns with totals <10. See spreadsheet for what I did about when to consider {l,ɹ,[+nas]} as pre-$C_2$ and when as $C_2$ itself, and some other tricky cases.

[We could get deeper into this: if $C_1$ is [+nasal], is there less likely to be a nasal preceding $C_2$? Similarly for liquids.]

Suppose English speakers have learned this pattern (see Coetzee 2010 for evidence that they do, at least for *s*CVC words; see Frisch & Zawaydeh 2001 for evidence that Arabic speakers know a similar but stronger pattern in Arabic).

- o How could the grammar express the pattern? What issues do we run into with GLA/MaxEnt, or with constraint indexing?

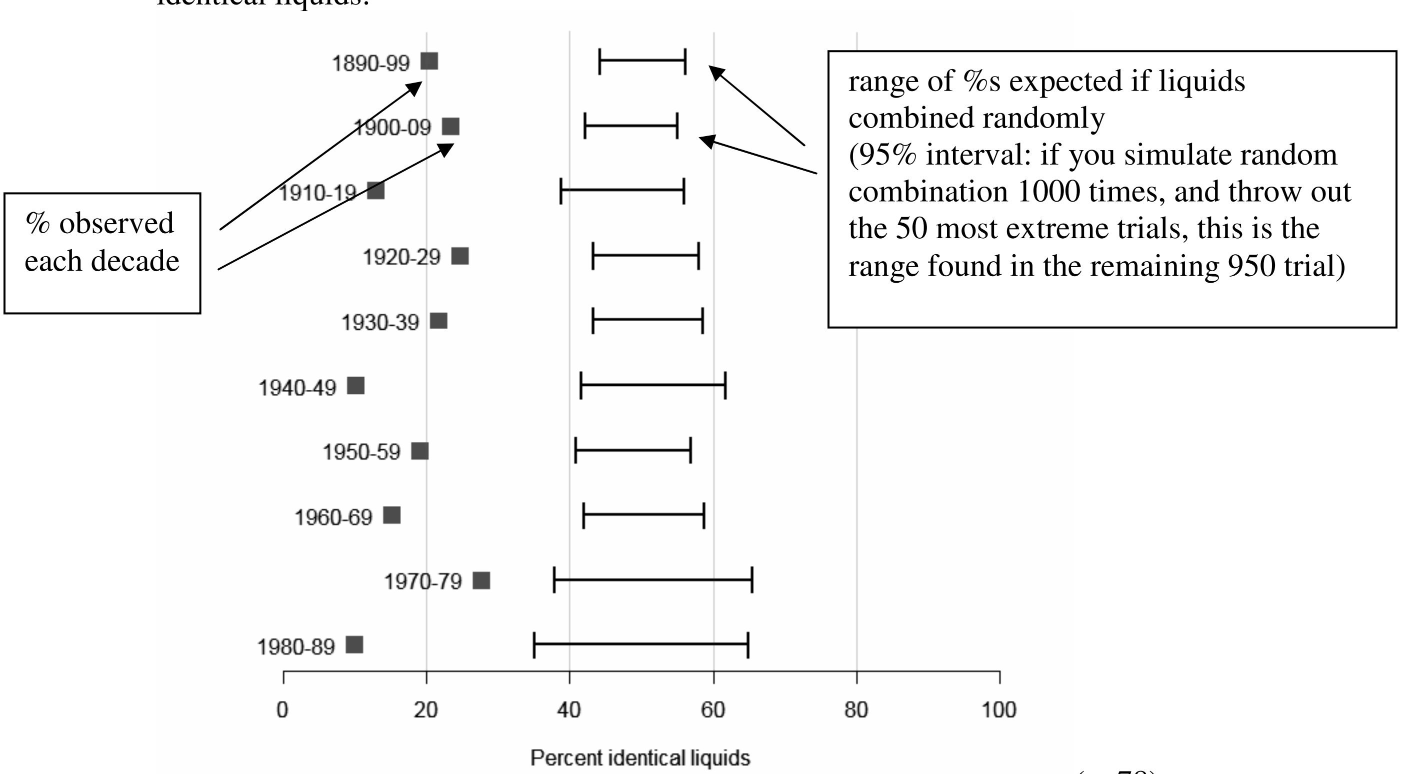
## 6. Lexical selection as an active shaper of the lexicon

In the English lexicon overall, if a word has two liquids, they're more likely to be *l...r* or *r...l* than *l...l* or *r...r*.
Martin 2007 shows... (pp. 76-77)

- In Old English, about 35% of words with two liquids have identical liquids, compared to ~55% expected by chance.
- In Middle English, it's about 25% (expect ~50%)
- Today, it's about 25% (expect ~50%)

Even though we've retained only ~10-15% of the Old English vocabulary!

Martin 2007 also uses the Oxford English Dictionary, which gives dates of earliest attested use for each word, to look at words newly entering the language. In every decade, new words avoid identical liquids:




(p. 78)

See Martin 2007 for an implemented model of lexical selection.

## 7. Filters in general

A lot of phonological and paraphonological activity has this character: it's not so much about mapping an input to an output as about deciding how good the resulting output is.

- Which new words enter the language, as we just saw
- Which names people choose for babies, fantasy role-playing characters, and pharmaceuticals (Martin 2007)
- Which first-name/last-name combinations people choose (Shih 2012)
- Which words make a good pun (Fleischhacker 2006)
- Which pairs of words make a good compound (Martin 2004; Martin 2007; Martin 2011)
- Which lines of poetry are legal (Hayes 2009)
- Which words can take which affixes (a big literature, but see Orgun & Sprouse 1999 in particular for the idea of a filter); we'll see more of this in weeks 5 and 8

See Martin 2007 for an implemented model of how different options compete. Options like...

- *couch* vs. *sofa*
- *carp pond* vs. *koi pond*
- *Mainer* vs. *Mainean* vs. *Maineite* (for 'person from the state of Maine')

pushing the idea further...

- writing a love song about the moon in June vs. writing one about the sun in August
- drawing a cartoon with the pun *Napoleon Blown-apart* in its caption vs. drawing a cartoon about something else, or not drawing a cartoon at all.

Crucially, one factor in the competition is how phonologically "good" a competitor is.
=> Even if each tableau has just one winner (/kawtʃ/ → [kawtʃ], /sowfʌ/ → [sowfʌ]), the grammar should still attach a goodness score to it (see Coetzee & Pater 2007 for a way to do it)

Going back to our English example: a hypothetical input /spowm/ can surface faithfully, but with a poor score attached to it, so that it's unlikely to catch on as a new word:

| /spowm/ | Max-C | Ident(place)/__{V,#} | *s[labial]...[labial] |
|---|---|---|---|
| ☞ *a* spowm | | | * |
| *b* spow | *! | | |
| *c* stowm | | *! | |
| *compare* | | | |
| /spown/ | Max-C | Ident(place)/__{V,#} | *s[labial]...[labial] |
| ☞ *d* spown | | | |
| *e* spow | *! | | |
| *f* stown | | *! | |

- Discuss ways to give different scores to (a) and (d) above

## 8. Wrapping up the week

- We now have two competing tools for handling free variation: Stochastic OT (+Gradual Learning Algorithm) and MaxEnt OT (+learning algorithms we won't worry about).
- We've discussed the problem of lexical variation: how to allow each word to prefer one variant, while still capturing in the grammar the distribution of the variants?
    - We saw constraint indexing as one approach
- We've discussed the related problem of lexical selection and filters in general: how to capture in the grammar the underrepresentation of certain word types?
    - We left this somewhat open, but considered attaching numbers to winning candidates that a higher-level selection process can use

**Next time:** The nuts and bolts of process application, in SPE and OT.
For example, suppose we have a rule i → Ø / VC__CV and the word /taminisiko/--what happens? Suppose the rule is optional—what then?

**Today's bottom line**
- You're now well equipped to model data with free variation (e.g., on a homework assignment!).
- We've dealt conceptually with lexical variation and lexical selection, but actually modeling them would require some extra thought (e.g., in a term paper)

**References**

Coetzee, Andries W. 2010. Gradient well-formedness in Harmonic Grammar: phonological performance as a window on phonological competence. *Journal of the Phonetic Society of Japan* 14. 13-23.

Coetzee, Andries W & Joe Pater. 2007. Weighted constraints and gradient phonotactics in Muna and Arabic.

Fleischhacker, Heidi. 2006. Similarity in phonology: evidence from reduplication and loan adaptation.. UCLA ph.d. dissertation.

Frisch, Stefan A, Janet B Pierrehumbert & Michael B Broe. 2004. Similarity Avoidance and the OCP. *Natural Language & Linguistic Theory* 22(1). 179–228.

Frisch, Stefan A. & Bushra Adnan Zawaydeh. 2001. The Psychological Reality of OCP-Place in Arabic. *Language* 77(1). 91-106. (7 January, 2012).

Goldwater, Sharon & Mark Johnson. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model.. In Jennifer Spenader, Anders Eriksson, & Östen Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111-120. Stockholm: Stockholm University.

Hayes, Bruce. 2009. Faithfulness and componentiality in metrics.. In Sharon Inkelas & Kristin Hanson (eds.), *The nature of the word*, 113-148. Cambridge, MA: MIT Press.

Martin, Andrew. 2004. The effects of distance on lexical bias: sibilant harmony in Navajo compounds.. UCLA master's thesis.

Martin, Andrew. 2007. The evolving lexicon.. University of California, Los Angeles ph.d. dissertation.

Martin, Andrew. 2011. Grammars leak: modeling how phonotactic generalizations interact within the grammar. *Language* 87(4). 751-770.

Orgun, Cemil Orhan & Ronald L Sprouse. 1999. From "MParse" to "Control": Deriving Ungrammaticality. *Phonology* 16(2). 191–224.

Pater, Joe. 2008. Gradual Learning and Convergence. *Linguistic Inquiry*.

Shewchuk, Jonathan Richard. 1994. An introduction to the Conjugate Gradient Method without the agonizing pain.. Manuscript. Carnegie Mellon University, ms.

Shih, Stephanie. 2012. Linguistic determinants in English personal name choice.. Presentation. Paper presented at the LSA annual meeting, Portland, OR.