The OpenLexicons Project - Development and Uses of SUBTLEX-Corpora for Investigating Sound Symbolism and Brazilian Portuguese.

Kevin Tang Paweł Mandera Emmanuel Keuleers
kevin.tang.10@ucl.ac.uk pawel.mandera@ugent.be
emmanuel.keuleers@ugent.be

Department of Linguistics, University College London
Department of Experimental Psychology, Ghent University

3rd NetWordS Workshop, Dubrovnik, 2013

crr.ugent.be

Outline

1. Introduction
2. Corpus Enrichment
   - Beyond Token Frequency
3. Lexicon Modelling
   - Corpus, Lemmatisation and Morphemisation
4. Analyses
   - Measures of phonetic similarity
   - Weighting schemes
   - Evaluation
5. Conclusion

SUBTLEX
Phonological & Psycholinguistic Research Tools

- SUBTLEX film subtitle frequencies are excellent predictors of behavioral task measures for English [Brysbaert and New, 2009], French [New et al., 2007], Dutch [Keuleers et al., 2010] ...
- These subtitles are mostly from English-language movies from all genres, and show a wide range of tenses, persons, speech act types in the dialogues.
- In this presentation, we demonstrate the richness of SUBTLEX beyond the token frequency norms, and subsequently use an enriched corpus to model aspects of the lexicon.

Corpus Enrichment	SUBTLEX Brazilian Portuguese
Lexicon Modelling	Sound Symbolism in English

SUBTLEX: Beyond Token Frequency

While most corpora stop at token frequency, we here focus on the possible enrichments. We demonstrate them on SUBTLEX-BR-PT, a 61mil Brazilian Portuguese corpus.

1. Pseudo Words
2. N-gram
3. Contextual Diversity
4. Grapheme to Phoneme Conversion
5. Lexical Neighbourhood Density
6. Lemmatisation and POS-Tagging

Generating Pseudo Words in a *Principled* Way

Pseudo-words play a crucial role in linguistic research, from testing morphophonological productivity to getting reaction times of words through lexicon decision tasks.

1. Change one letter/phoneme from a real word, e.g. milk – *pilk, malk, mirk*.... Such was used in the English Lexicon Project [Balota et al., 2007]
2. ARC nonword database [Rastle et al., 2002] – *Monosyllabic* only
3. Stringing together high-frequency bigrams or trigrams. WordGen [Duyck et al., 2004] – *Slow* with long words, more likely to have phonotactic-illegality

Wuggy: a multilingual pseudoword generator

Wuggy [Keuleers and Brysbaert, 2010]

- ✓ Multilingual (Alphabetic languages)
- ✓ Perfect for mega studies (Extremely quick)
- ✓ Simple to use and implement (Transparent Python codes)
- ✓ Legal phonotactics

- Currently makes pseudowords in Basque, Dutch, English, French, German, Serbian (Cyrillic and Latin), Spanish and Vietnamese
- Requires only a syllabified word list (orthography) and a list of possible orthographical nuclei.

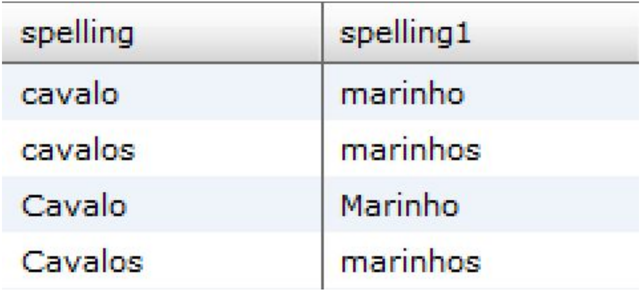
Brazilian Portuguese Wugs

- Brazilian Portuguese Module (In progress, not yet available online)
- The Subtlex-Br-Pt Word List was used.
- Brazilian Portuguese syllabification was performed using Lingua-PT-Hyphenate Perl Module by José Alves de Castro

Beyond Unigram – Bigram

- Bigram word corpus would allow searching of potential compounds and collocation frequency.

Cavalo-marinho “Seahorse”



spelling	spelling1
cavalo	marinho
cavalos	marinhos
Cavalo	Marinho
Cavalos	marinhos

Contextual Diversity

- Contextual diversity (CD) is a measure of the number of document/context files that a word has occurred in (in our case, subtitle files)
- CD could be better than token frequency in capturing word-naming and lexical decision times in terms of capturing more variances [Adelman et al., 2006, Brysbaert and New, 2009]
- This has not been widely used in linguistics which currently prefers the use of token frequency [Bybee, 1995, 2003, Huback, 2007, Coetzee and Kawahara, in press, 2013]

Grapheme to Phone Conversion

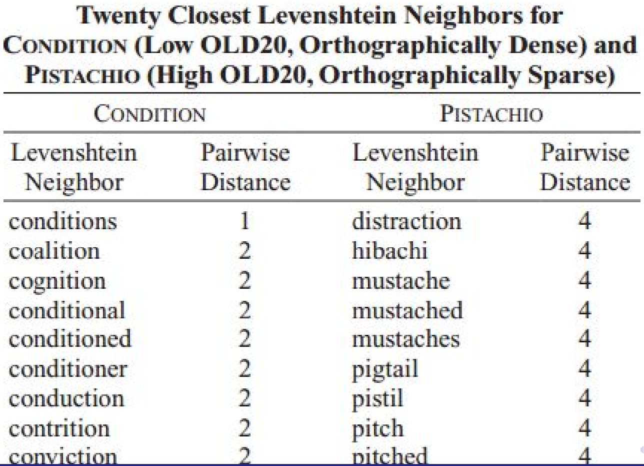
- Algorithm-based converter– Hard-coded rules to map graphemes to phones
- Probabilistic models – Train on pronunciation dictionaries
- No readily available converter for Brazilian Portuguese, so a European Portuguese converter was used, with added hard-coded rules (in progress).
  `http://www.co.it.pt/~labfala/g2p/`
  (Signal Processing Lab, Instituto de Telecomunicações)

Lexical Neighbourhood Density

- Why bother? See Luce and Pisoni [1998]
- Orthographical and Phonological
- One edit distance metric
- Coltheart's N (the number of words that are one substitution away)
- **Orthographic Levenshtein distance 20 (OLD20)**

Orthographic Levenshtein distance 20

- Average Levenshtein distance of the 20 closest neighbours.
- Suggested to be a better metric than Colheart's N in predicting performance in behavioural tasks [Yarkoni et al., 2008]

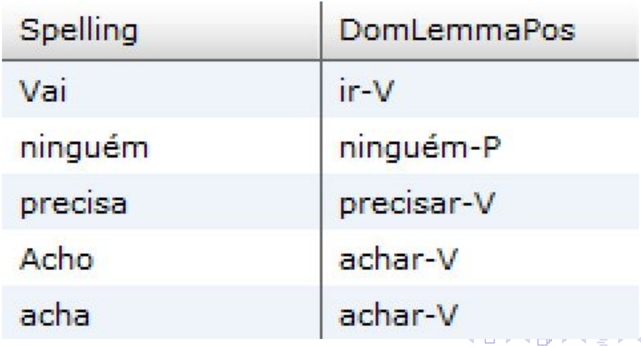


Twenty Closest Levenshtein Neighbors for CONDITION (Low OLD20, Orthographically Dense) and PISTACHIO (High OLD20, Orthographically Sparse)

CONDITION		PISTACHIO	
Levenshtein Neighbor	Pairwise Distance	Levenshtein Neighbor	Pairwise Distance
conditions	1	distraction	4
coalition	2	hibachi	4
cognition	2	mustache	4
conditional	2	mustached	4
conditioned	2	mustaches	4
conditioner	2	pigtail	4
conduction	2	pistil	4
contrition	2	pitch	4
conviction	2	pitched	4

Lemmatisation and POS – Tagging

- Determining the lemma and Part of Speech for a given word
- e.g. Lemma {‘walk’} – Form {‘walk’, ‘walked’, ‘walks’, ‘walking’}
- TreeTagger for Portuguese by Pablo Gamallo was used `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`



Spelling	DomLemmaPos
Vai	ir-V
ninguém	ninguém-P
precisa	precisar-V
Acho	achar-V
acha	achar-V

Corpus URL

Different versions of the corpus (with different filters) with an interactive interface are available at
`http://crr.ugent.be/subtlex-pt-br/`

For more specific corpora:
Unigram:
`http://zipf.ugent.be/open-lexicons/interfaces/pb-subtitles-unigram/`
Bigram:
`http://zipf.ugent.be/open-lexicons/interfaces/br-pt-bigrams/`
Lemmatised + POS-Tagged:
`http://zipf.ugent.be/open-lexicons/interfaces/br-pt-lemmas/`

Modelling Sound Symbolism

- With an enriched SUBTLEX corpus, we are now ready to model aspects of the lexicon.
- Sound symbolism [Sapir, 1929]
- Whether the link between sound and meaning is arbitrary?
- An important way human languages innovate lexical items

> *"In general, linguistic theory assumes that the relation between sound and meaning is arbitrary. Any aspect of language that goes against this assumption has traditionally been considered as only a minor exception to the general rule." [Hinton et al., 2006, Ch.1, p.1]*

Sound Symbolism – a New Visit to an Old Topic

- Comparing basic vocabulary cross-linguistically [Wichmann et al., 2010]
- Testing the perception of phonetic properties [Sapir, 1929, Newman, 1933] e.g. [a](*"large"*) versus [i](*"small"*)
- Validating phonesthemes [Householder, 1946, Drellishak, 2006] e.g. English *'gl'* – "light"-related.

Our Approach

- Reconstruction of Meaning from Sound
- SUBTLEX English Corpus
- Topic Modelling

Corpus, Lemmatisation and Morphemisation

- Subtitle-corpus containing 69,382 files and 385 mil. tokens
- The corpus was tagged and lemmatized using Stanford tagger [Toutanova et al., 2003] because the inflected forms of a lemma will have similar semantic content as well as phonetic content, e.g. *laugh-ing* and *laugh-ed*
- Lemmas broken into morphemes using CELEX [Baayen et al., 1995] e.g. *unnecessarily* would be broken down into three morphemes *un*, *necessary*, and *ly*

Semantic space

- Latent Dirichlet Allocation [Blei et al., 2003] - a simple topic modeling technique was shown to outperform LSA [Landauer and Dumais, 1997] in predicting human associations [Griffiths et al., 2007]
- Each topic represented as a probability distribution over words
- Each document represented as a probability distribution over topics
- The morphemized corpus was used to train different topic models (400,1200 topics)

Example topics

Topic	Key words
1	eat rice soup bean look food hot noodle day bowl buy water
2	car engine drive fast speed ly tank look mile er gear gas
3	minister ment govern ion prime ly politic ambassador
4	plane air fly flight pilot land crash port jet craft
5	bomb ion blow explode hostage time move explode ion ive
6	priest church god father saint bishop holy pope ion confess
7	majesty emperor prince ness palace royal ly excellency
…	…

Analyses

- As a first step, used only *monomorphemic* and *monosyllabic* words.
- n.b. CELEX is *extremely* conservative about monomorphemicity
- 3248 morphemes

Measures of phonetic similarity I

Three different distance/similarity metrics were explored on Orthographical and Phonemic forms.

Segmental

- Levenshtein distance [Levenshtein, 1966] (Ortho/Phon)
- Blind to featural differences: Manner, Place of Articulation and Voicing
- e.g. LD of 1: /**p**ɪt/–/**g**ɪt/, /**b**ɪt/–/**g**ɪt/

Featural

- ALINE [Kondrak, 2002, Huff, 2010] (Phon)
- Phonetic features. Locally aligned to detect phonesthemes

Measures of phonetic similarity II

- Subsyllabic Modified Value Difference Metric (henceforth Sub-MVDM) (Ortho) [Cost and Salzberg, 1993, Keuleers and Daelemans, 2007]:
- Calculates similarity matrix for each subsyllabic segment (with the subsyllabic segment as a feature and a pair of the two remaining segments as a class)
- $dist(w_1, w_2) = dist(onset_{w_1}, onset_{w_2}) + dist(nucleus_{w_1}, nucleus_{w_2}) + dist(coda_{w_1}, coda_{w_2})$
- $dist(onset_1, onset_2) = \sum_{n \in (nucleus, coda)} |P(n|onset_1) - P(n|onset_2)|$

Reconstruct Meaning from Sound

- Leave-one-out method
- Reconstruct the semantic vector for each word using *only* the semantic vectors of the remaining words
- **Weighted** by their corresponding phonetic similarity with the words that are being reconstructed.

Weighting schemes

None
: No weighting schemes, directly use the phonetic similarity between each word and the remaining words.

Shepard
: Apply a weighting decay function [Shepard et al., 1987], alpha and beta parameters, to the phonetic similarity

N-th neighbours
: Use only the phonetic similarity of the n-th closest neighbours [Luce and Pisoni, 1998, Yarkoni et al., 2008].

Evaluation

1. Above random
2. Reconstructability

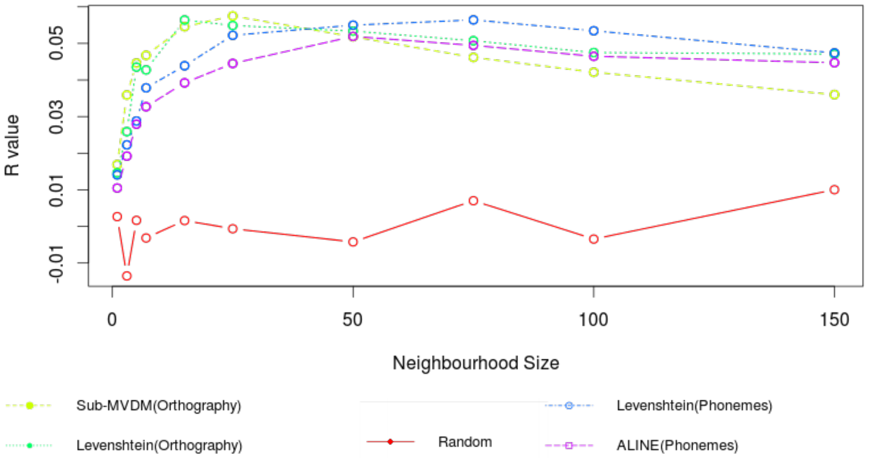
Above random

- Compared the relative semantic similarity between original words and reconstructed words
- Varied the number of neighbours, metrics and topic sizes
- Calculated the correlation value between the original semantic space, and two reconstructed semantic space (weighted by phonetic similarity or by random values)

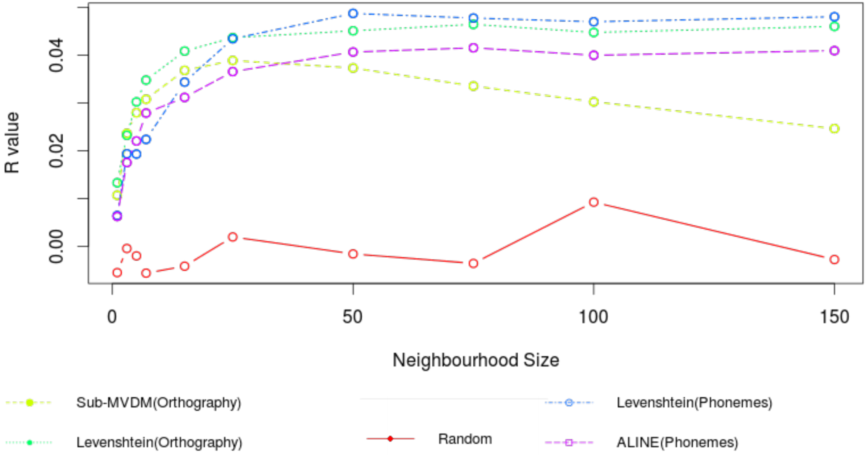
Our Question

Can the semantic space be reconstructed phonetically *above random*?

Above random: 400 Topics



Above random: 1200 Topics



Above random? *Yes, it is*

- Consistently above random across all metrics and all topic sizes.
- Effect of locality: Optimal neighbourhood size begins ≈ 20
- cf. OLD20 – *“The increment in variance explained ... peaked around 10–20 words ...”* [Yarkoni et al., 2008]
- Orthography reaches its peak quicker that Phonemes, i.e. with fewer neighbours.
- Why Orthography outperformed Phonemes?
  English spelling does not generally reflect the sound changes in the pronunciation. Consider ni**gh**t–lau**gh**, **g**naw, lam**b**

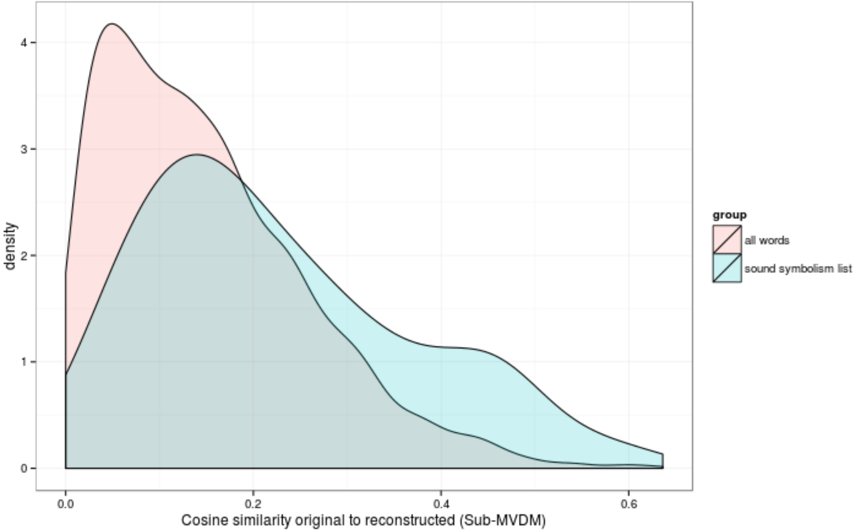
Reconstructability

- Many analyses of sound symbolism were based on words which are classified as being symbolically motivated.
- The classification of these words can be highly subjective
- e.g. *'gl'* in English: commonly found in "light"-related words.
- *'gleam'*, *'glow'*, but what about *'glad'*?
- $\approx$ 120 sound symbolic words were extracted from Hinton et al. [2006][Ch.19] for comparison

Our Question

Are these allegedly sound-symbolic words more reconstructable than just any word?

Reconstructability



Reconstructability

- The sound symbolic words are *skewed* in the positive direction of reconstructablilty.
- Welch Two Sample t-test showed that the distributions of reconstructability of all the words and of the sound symbolic words are significantly different, *p*-value = 2.12e-08
- Suggesting that these allegedly sound symbolic words are more reconstructable than any words.

Conclusion

- The existence of sound symbolism in the English lexicon is small, but the link between sound and meaning is ***Not* Arbitrary**, as it can be reconstructed above random.
- There’s a clear **Locality** effect of neighbours
- Using **Reconstructability**, the results confirmed native speakers’ intuitions of sound symbolism, that is, sound-symbolically motivated words are more reconstructable than other words.

Conclusion

- Our creation of very large SUBTLEX corpora, openly available and in a standardized format, will remain accessible as a potentially **Valuable Resource** for Phonological & Psycholinguistic Research and for a number of adjacent fields.
- We demonstrated the **Richness** of SUBTLEX corpora and their potential for research beyond token frequency, ranging from **Pseudo-word Creation** with Wuggy to **Lexicon Modelling** for Sound Symbolism.

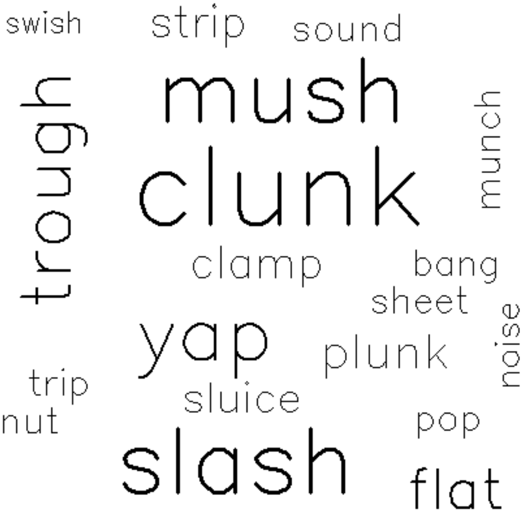
Acknowledgement

We would like to thank the following researchers from the Center for Reading Research (CRR), Ghent University, Belgium

- Prof. Marc Brysbaert

European Science Foundation - Short Visit Grant received by NetWordS - 09-RNP-089

Reconstructablilty



References I

J.S. Adelman, G.D.A. Brown, and J.F. Quesada. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823, 2006.

H.R. Baayen, R. Piepenbrock, and L. Gulikers. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, 1995.

D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman. The english lexicon project. *Behavior Research Methods*, 39(3):445–459, 2007.

D.M Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

M. Brysbaert and B. New. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4): 977–990, Nov 2009.

J. Bybee. Regular morphology and the lexicon. *Language and cognitive processes*, 10(5):425–455, 1995.

References II

J. Bybee. *Phonology and language use*, volume 94. Cambridge University Press, 2003.

A.W. Coetzee and S. Kawahara. Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, 31(1), in press, 2013.

S. Cost and S. Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine learning*, 10(1):57–78, 1993.

S. Drellishak. Statistical techniques for detecting and validating phonesthemes. *Unpublished masters thesis, University of Washington*, 2006.

W. Duyck, T. Desmet, L.P.C. Verbeke, and M. Brysbaert. Wordgen: A tool for word selection and nonword generation in dutch, english, german, and french. *Behavior Research Methods, Instruments, & Computers*, 36(3):488–499, 2004.

T.L. Griffiths, M. Steyvers, and J.B. Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

L. Hinton, J. Nichols, and J.J. Ohala. *Sound symbolism*. Cambridge University Press, 2006.

F.W. Householder. On the problem of sound and meaning, an english phonestheme. *Word*, 2:83, 1946.

References III

A. Huback. *Efeitos de freqüência nas representações mentais*. PhD thesis, Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

P. Huff. Pyaline: Automatically growing language family trees using the aline distance. Master's thesis, Brigham Young University, 2010.

E. Keuleers and M. Brysbaert. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633, 2010.

E. Keuleers and W. Daelemans. Memory-based learning models of inflectional morphology: A methodological case-study. *Lingue e linguaggio*, 6(2):151–174, 2007.

E. Keuleers, M. Brysbaert, and B. New. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3):643–650, Aug 2010.

G. Kondrak. *Algorithms for Language Reconstruction*. Ph.D. dissertation, University of Toronto, 2002.

T.K. Landauer and S.T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

References IV

V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

P.A. Luce and D.B. Pisoni. Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1):1, 1998.

B. New, M. Brysbaert, J. Veronis, and C. Pallier. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4):661–677, 2007.

S.S. Newman. Further experiments in phonetic symbolism. *The American Journal of Psychology*, 45(1):53–75, 1933.

K. Rastle, J. Harrington, and M. Coltheart. 358,534 nonwords: The arc nonword database. *The Quarterly Journal of Experimental Psychology: Section A*, 55(4): 1339–1362, 2002.

E. Sapir. A study in phonetic symbolism. *Journal of experimental Psychology*, 12(3): 225, 1929.

R.N Shepard et al. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

References V

K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073478. URL http://dx.doi.org/10.3115/1073445.1073478.

S. Wichmann, E.W. Holman, and C.H. Brown. Sound symbolism in basic vocabulary. *Entropy*, 12(4):844–858, 2010.

T. Yarkoni, D. Balota, and M. Yap. Moving beyond Colthearts N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979, 2008.