Biases in Predicting the Human Language Model

Alex B. Fine
University of Illinois at Urbana-Champaign
abfine@illinois.edu

Austin F. Frank
Riot Games
aufrank@riotgames.com

T. Florian Jaeger
University of Rochester
fjaeger@bcs.rochester.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@cs.jhu.edu

Abstract

We consider the prediction of three human behavioral measures – lexical decision, word naming, and picture naming – through the lens of domain bias in language modeling. Contrasting the predictive ability of statistics derived from 6 different corpora, we find intuitive results showing that, e.g., a British corpus overpredicts the speed with which an American will react to the words *ward* and *duke*, and that the Google n-grams overpredicts familiarity with technology terms. This study aims to provoke increased consideration of the human language model by NLP practitioners: biases are not limited to differences between corpora (i.e. "train" vs. "test"); they can exist as well between corpora and the intended user of the resultant technology.

1 Introduction

Computational linguists build statistical language models for aiding in natural language processing (NLP) tasks. Computational psycholinguists build such models to aid in their study of human language processing. Errors in NLP are measured with tools like precision and recall, while errors in psycholinguistics are defined as failures to model a target phenomenon.

In the current study, we exploit errors of the latter variety—failure of a language model to predict human performance—to investigate *bias* across several frequently used corpora in computational linguistics. The human data is revealing because it trades on the fact that human language processing is *probability-sensitive*: language processing reflects implicit knowledge of probabilities computed over linguistic units (e.g., words). For example, the amount of time required to read a word varies as a function of how predictable that word is (McDonald and Shillcock, 2003). Thus, failure of a language model to predict human performance reveals a mismatch between the language model and the human language model, i.e., bias.

Psycholinguists have known for some time that the ability of a corpus to explain behavior depends on properties of the corpus and the subjects (cf. Balota et al. (2004)). We extend that line of work by directly analyzing and quantifying this bias, and by linking the results to methodological concerns in both NLP and psycholinguistics.

Specifically, we predict human data from three widely used psycholinguistic experimental paradigms—lexical decision, word naming, and picture naming—using unigram frequency estimates from Google n-grams (Brants and Franz, 2006), Switchboard (Godfrey et al., 1992), spoken and written English portions of CELEX (Baayen et al., 1995), and spoken and written portions of the British National Corpus (BNC Consortium, 2007). While we find comparable overall fits of the behavioral data from all corpora under consideration, our analyses also reveal specific domain biases. For example, Google n-grams overestimates the ease with which humans will process words related to the web (*tech*, *code*, *search*, *site*), while the Switchboard corpus—a collection of informal telephone conversations between strangers—overestimates how quickly humans will react to colloquialisms (*heck, darn*) and backchannels (*wow, right*).

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 7–12,
Baltimore, Maryland, USA, June 23-25 2014.

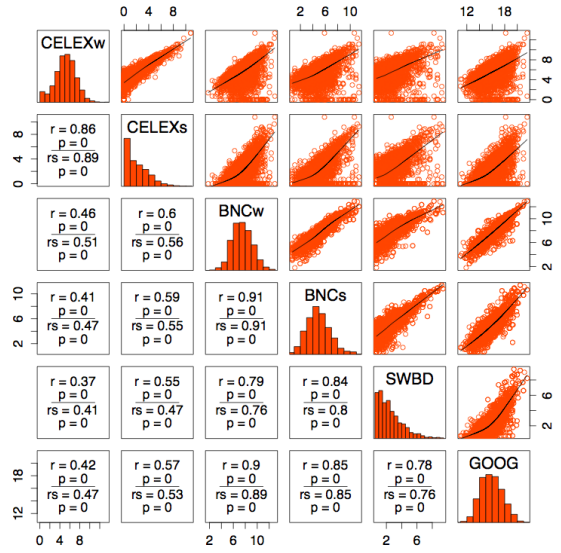

Figure 1: Pairwise correlations between log frequency estimates from each corpus. Histograms show distribution over frequency values from each corpus. Lower left panels give Pearson (top) and Spearman (bottom) correlation coefficients and associated p-values for each pair. Upper right panels plot correlations

2 Fitting Behavioral Data

2.1 Data

Pairwise Pearson correlation coefficients for log frequency were computed for all corpora under consideration. Significant correlations were found between log frequency estimates for all pairs (Figure 1). Intuitive biases are apparent in the correlations, e.g.: BNCw correlates heavily with BNCs (0.91), but less with SWBD (0.79), while BNCs correlates more with SWBD (0.84).[1]

Corpus	Size (tokens)
Google n-grams (web release)	$\sim$ 1 trillion
British National Corpus (written, BNCw)	$\sim$ 90 million
British National Corpus (spoken, BNCs)	$\sim$ 10 million
CELEX (written, CELEXw)	$\sim$ 16.6 million
CELEX (spoken, CELEXs)	$\sim$ 1.3 million
Switchboard (Penn Treebank subset 3)	$\sim$ 800,000

Table 1: Summary of the corpora under consideration.

2.2 Approach

We ask whether domain biases manifest as systematic errors in predicting human behavior. Log unigram frequency estimates were derived from each corpus and used to predict reaction times (RTs) from three experiments employing *lexical decision* (time required by subjects to correctly identify a string of letters as a word of English (Balota et al., 1999)); *word naming* (time required to read aloud a visually presented word (Spieler and Balota, 1997); (Balota and Spieler, 1998)); and *picture naming* (time required to say a picture's name (Bates et al., 2003)). Previous work has shown that more frequent words lead to faster RTs. These three measures provide a strong test for the biases present in these corpora, as they span written and spoken lexical comprehension and production.

To compare the predictive strength of log frequency estimates from each corpus, we fit mixed effects regression models to the data from each experiment. As controls, all models included (1) mean log bigram frequency for each word, (2) word category (noun, verb, etc.), (3) log morphological family size (number of inflectional and derivational morphological family members), (4) number of synonyms, and (5) the first principal component of a host of orthographic and phonological features capturing neighborhood effects (type and token counts of orthographic and phonological neighbors as well as forward and backward inconsistent words; (Baayen et al., 2006)). Models of lexical decision and word naming included random intercepts of participant age to adjust for differences in mean RTs between old (mean age = 72) vs. young (mean age = 23) subjects, given differences between younger vs. older adults' processing speed (cf. (Ramscar et al., 2014)). (All participants in the picture naming study were college students.)

2.3 Results

For each of the six panels corresponding to frequency estimates from a corpus A, Figure 2 gives the χ^2 value resulting from the log-likelihood ratio of (1) a model containing A and an estimate from one of the five remaining corpora (given on the x axis) and (2) a model containing just the corpus indicated on the x axis. Thus, for each panel, each bar in Figure 2 shows the explanatory power of estimates from the corpus given at the top of the panel after controlling for estimates from each of the other corpora.

Model fits reveal intuitive, previously undocumented biases in the ability of each corpus to predict human data. For example, corpora of British English tend to explain relatively little after con-

[1] BNCw and BNCs are both British, while BNCs and SWBD are both spoken.

trolling for other British corpora in modeling lexical decision RTs (yellow). Similarly, Switchboard provides relatively little explanatory power over the other corpora in predicting picture naming RTs (blue bars), possibly because highly imageable nouns and verbs frequent in everyday interactions are underrepresented in telephone conversations between people with no common visual experience. In other words, idiosyncratic facts about the topics, dialects, etc. represented in each corpus lead to systematic patterns in how well each corpus can predict human data relative to the others. In some cases, the predictive value of one corpus after controlling for another—apparently for reasons related to genre, dialect—can be quite large (cf. the χ^2 difference between a model with both Google and Switchboard frequency estimates compared to one with only Switchboard [top right yellow bar]).

In addition to comparing the overall predictive power of the corpora, we examined the words for which behavioral predictions derived from the corpora deviated most from the observed behavior (word frequencies strongly over- or under-estimated by each corpora). First, in Table 2 we give the ten words with the greatest relative difference in frequency for each corpus pair. For example, *fife* is deemed more frequent according to the BNC than to Google.[2]

These results suggest that particular corpora may be genre-biased in systematic ways. For instance, Google appears to be biased towards terminology dealing with adult material and technology. Similarly, BNCw is biased, relative to Google, towards Britishisms. For these words in the BNC and Google, we examined errors in predicted lexical decision times. Figure 3 plots errors in the linear model's prediction of RTs for older (top) and younger (bottom) subjects.

The figure shows a positive correlation between how large the difference is between the lexical decision RT predicted by the model and the actually observed RT, and how over-estimated the log frequency of that word is in the BNC relative to Google (left panel) or in Google relative to the BNC (right panel). The left panel shows that BNC produces a much greater estimate of the log frequency of the word *lee* relative to Google, which leads the model to predict a lower RT for this word than is observed (i.e., the error is positive; though note that the error is less severe for older relative to younger subjects). By contrast, the asymmetry between the two corpora in the estimated frequency of *sir* is less severe, so the observed RT deviates less from the predicted RT. In the right panel, we see that Google assigns a much greater estimate of log frequency to the word *tech* than the BNC, which leads a model predicting RTs from Google-derived frequency estimates to predict a far lower RT for this word than observed.

3 Discussion

Researchers in computational linguistics often assume that more data is always better than less data (Banko and Brill, 2001). This is true insofar as larger corpora allow computational linguists to generate *less noisy* estimates of the average language experience of the users of computational linguistics applications. However, corpus size does not necessarily eliminate certain types of *biases* in estimates of human linguistic experience, as demonstrated in Figure 3.

Our analyses reveal that 6 commonly used corpora fail to reflect the human language model in various ways related to dialect, modality, and other properties of each corpus. Our results point to a type of bias in commonly used language models that has been previously overlooked. This bias may limit the effectiveness of NLP algorithms intended to generalize to a linguistic domains whose statistical properties are generated by humans.

For psycholinguists these results support an important methodological point: while each corpus presents systematic biases in how well it predicts human behavior, all six corpora are, on the whole, of comparable predictive value and, specifically, the results suggest that the web performs as well as traditional instruments in predicting behavior. This has two implications for psycholinguistic research. First, as argued by researchers such as Lew (2009), given the size of the Web compared to other corpora, research focusing on low-frequency linguistic events—or requiring knowledge of the distributional characteristics of varied contexts—is now more tractable. Second, the viability of the web in predicting behavior opens up possibilities for computational psycholinguistic research in languages for which no corpora exist (i.e., most

[2] Surprisingly, *fife* was determined to be one of the words with the largest frequency asymmetry between Switchboard and the Google n-grams corpus. This was a result of lowercasing all of the words in in the analyses, and the fact that Barney Fife was mentioned several times in the BNC.

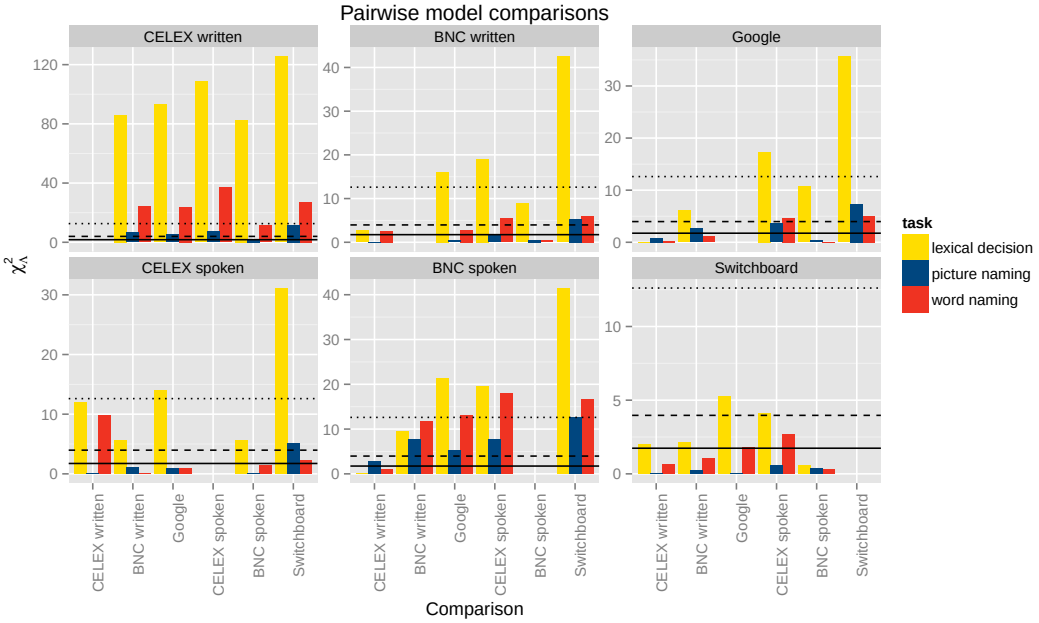

Figure 2: Results of log likelihood ratio model comparisons. Large values indicate that the reference predictor (panel title) explained a large amount of variance over and above the predictor given on the x-axis.

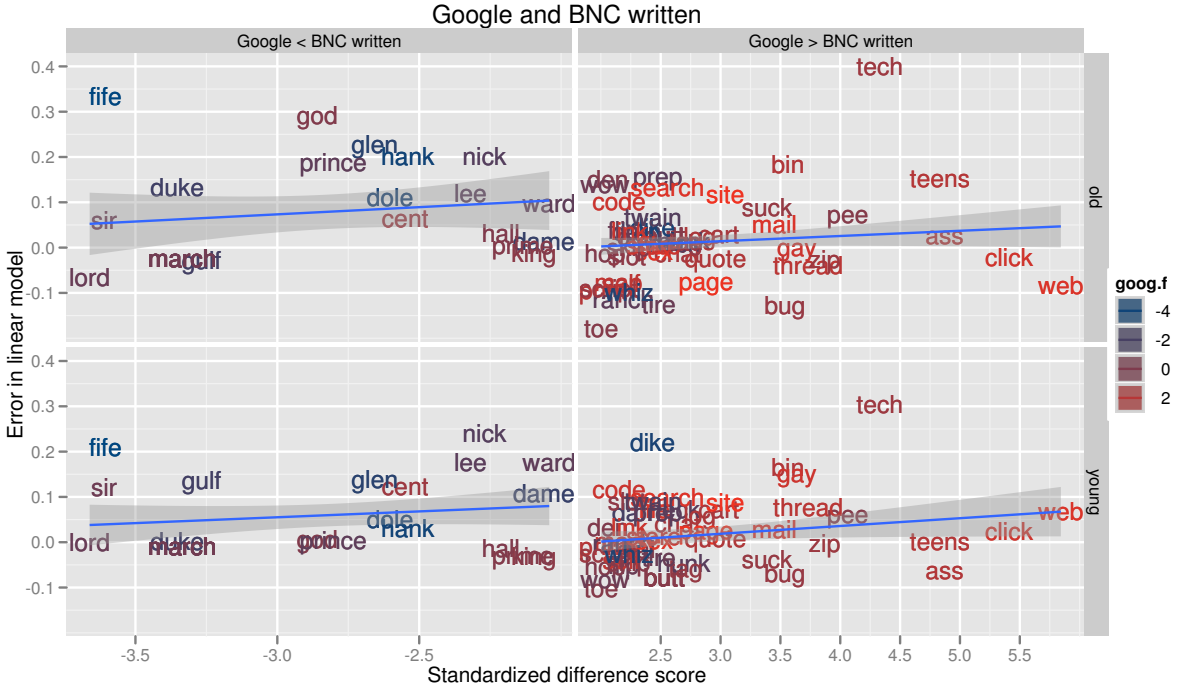

Figure 3: Errors in the linear model predicting lexical decision RTs from log frequency are plotted against the standardized difference in log frequency in the Google n-grams corpus versus the written portion of the BNC. Top and bottom panels show errors for older and younger subjects, respectively. The left panel plots words with much greater frequency in the written portion of the BNC relative to Google; the right panel plots words occurring more frequently in Google. Errors in the linear model are plotted against the standardized difference in log frequency across the corpora, and word color encodes the degree to which each word is more (red) or less (blue) frequent in Google. That the fit line in each graph is above 0 in the y-axis means that on average these biased words in each domain are being over-predicted, i.e., the corpus frequencies suggest humans will react (sometimes much) faster than they actually did in the lab.

Greater	Lesser	Top-10
google	bnc.s	web, ass, gay, tire, text, tool, code, woe, site, zip
google	bnc.w	ass, teens, tech, gay, bug, suck, site, cart, log, search
google	celex.s	teens, cart, gay, zip, mail, bin, tech, click, pee, site
google	celex.w	web, full, gay, bin, mail, zip, site, sake, ass, log
google	swbd	gay, thread, text, search, site, link, teens, seek, post, sex
bnc.w	google	fife, lord, duke, march, dole, god, cent, nick, dame, draught
bnc.w	bnc.s	pact, corps, foe, tract, hike, ridge, dine, crest, aide, whim
bnc.w	celex.s	staff, nick, full, waist, ham, lap, knit, sheer, bail, march
bnc.w	celex.w	staff, lord, last, nick, fair, glen, low, march, should, west
bnc.w	swbd	rose, prince, seek, cent, text, clause, keen, breach, soul, rise
celex.s	google	art, yes, pound, spoke, think, mean, say, thing, go, drove
celex.s	bnc.s	art, hike, pact, howl, ski, corps, peer, spoke, jazz, are
celex.s	bnc.w	art, yes, dike, think, thing, sort, mean, write, pound, lot
celex.s	celex.w	yes, sort, thank, think, jazz, heck, tape, well, fife, get
celex.s	swbd	art, cell, rose, spoke, aim, seek, shall, seed, text, knight
celex.w	google	art, plod, pound, shake, spoke, dine, howl, sit, say, draught
celex.w	bnc.s	hunch, stare, strife, hike, woe, aide, rout, yell, glaze, flee
celex.w	bnc.w	dike, whiz, dine, shake, grind, jerk, whoop, say, are, cram
celex.w	celex.s	wrist, pill, lawn, clutch, stare, spray, jar, shark, plead, horn
celex.w	swbd	art, rose, seek, aim, rise, burst, seed, cheek, grin, lip
swbd	google	mow, kind, lot, think, fife, corps, right, cook, sort, do
swbd	bnc.s	creek, mow, guess, pact, strife, tract, hank, howl, foe, nap
swbd	bnc.w	stuff, whiz, tech, lot, kind, creek, darn, dike, bet, kid
swbd	celex.s	wow, sauce, mall, deck, full, spray, flute, rib, guy, bunch
swbd	celex.w	heck, guess, right, full, stuff, lot, last, well, guy, fair

Table 2: Examples of words with largest difference in z-transformed log frequencies (e.g., the relative frequencies of *fife*, *lord*, and *duke*, in the BNC are far greater than in Google).

languages). This furthers the arguments of the "the web as corpus" community (Kilgarriff and Grefenstette, 2003) with respect to psycholinguistics.

Finally, combining multiple sources of frequency estimates is one way researchers may be able to reduce the prediction bias from any single corpus. This relates to work in automatically building domain specific corpora (e.g., Moore and Lewis (2010), Axelrod et al. (2011), Daumé III and Jagarlamudi (2011), Wang et al. (2014), Gao et al. (2002), and Lin et al. (1997)). Those efforts focus on building representative document collections for a target domain, usually based on a seed set of initial documents. Our results prompt the question: can one use human behavior as the *target* in the construction of such a corpus? Concretely, can we build corpora by optimizing an objective measure that minimizes error in predicting human reaction times? Prior work in building balanced corpora used either rough estimates of the ratio of genre styles a normal human is exposed to daily (e.g., the Brown corpus (Kucera and Francis, 1967)), or simply sampled text evenly across genres (e.g., COCA: the Corpus of Contemporary American English (Davies, 2009)). Just as language models have been used to predict reading grade-level of documents (Collins-Thompson and Callan, 2004), human language models could be used to predict the appropriateness of a document for inclusion in an "automatically balanced" corpus.

4 Conclusion

We have shown intuitive, domain-specific biases in the prediction of human behavioral measures via corpora of various genres. While some psycholinguists have previously acknowledged that different corpora carry different predictive power, this is the first work to our knowledge to systematically document these biases across a range of corpora, and to relate these predictive errors to domain bias, a pressing issue in the NLP community. With these results in hand, future work may now consider the automatic construction of a "properly" balanced text collection, such as originally desired by the creators of the Brown corpus.

Acknowledgments

The authors wish to thank three anonymous ACL reviewers for helpful feedback. This research was supported by a DARPA award (FA8750-13-2-0017) and NSF grant IIS-0916599 to BVD, NSF IIS-1150028 CAREER Award and Alfred P. Sloan Fellowship to TFJ, and an NSF Graduate Research Fellowship to ABF.

References

A. Axelrod, X. He, and J. Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 11)*.

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX Lexical Database (Release 2). Linguistic Data Consortium, Philadelphia.

R. H. Baayen, L. F. Feldman, and R. Schreuder. 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53:496–512.

D. A. Balota and D. H. Spieler. 1998. The utility of item-level analyses in model evaluation: A reply to Seidenberg & Plaut (1998). *Psychological Science*.

D. A. Balota, M. J. Cortese, and M. Pilotti. 1999. Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society*, page 44.

D. Balota, M. Cortese, S. Sergent-Marshall, D. Spieler, and M. Yap. 2004. Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, (133):283316.

M. Banko and E. Brill. 2001. Mitigating the paucity of data problem. *Human Language Technology*.

E. Bates, S. D'Amico, T. Jacobsen, A. Szkely, E. Andonova, A. Devescovi, D. Herron, CC Lu, T. Pechmann, C. Plh, N. Wicha, K. Federmeier, I. Gerdjikova, G. Gutierrez, D. Hung, J. Hsu, G. Iyer, K. Kohnert, T. Mehotcheva, A. Orozco-Figueroa, A. Tzeng, and O. Tzeng. 2003. Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, 10(2):344–380.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

T. Brants and A. Franz. 2006. Web 1T 5-gram Version 1. Linguistic Data Consortium (LDC).

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200.

H. Daumé III and J. Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 11)*.

M. Davies. 2009. The 385+ million word corpus of contemporary american english (19902008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

J. Gao, J. Goodman, M. Li, and K. F. Lee. 2002. Toward a unified approach to statistical language modeling for chinese. In *Proceedings of the ACM Transactions on Asian Language Information Processing (TALIP 02)*.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP-92*, pages 517–520.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

H. Kucera and W.N. Francis. 1967. Computational analysis of present-day american english. providence, ri: Brown university press.

R. Lew, 2009. *Contemporary Corpus Linguistics*, chapter The Web as corpus versus traditional corpora: Their relative utility for linguists and language learners, pages 289–300. London/New York: Continuum.

S. C. Lin, C. L. Tsai, L. F. Chien, K. J. Chen, and L. S. Lee. 1997. Chinese language model adaptation based on document classification and multiple domain-specific language models. In *Proceedings of the 5th European Conference on Speech Communication and Technology*.

S.A. McDonald and R.C. Shillcock. 2003. Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological science*, 14(6):648–52, November.

R. C. Moore and W. Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 10)*.

M. Ramscar, P. Hendrix, C. Shaoul, P. Milin, and R. H. Baayen. 2014. The myth of cognitive decline: nonlinear dynamics of lifelong learning. *Topics in Cognitive Science*, 32:5–42.

D. H. Spieler and D. A. Balota. 1997. Bringing computational models of word naming down to the item level. 6:411–416.

L. Wang, D.F. Wong, L.S. Chao, Y. Lu, and J. Xing. 2014. A systematic comparison of data selection criteria for smt domain adaptation. *The Scientific World Journal*.