Target word prediction and paraphasia classification in spoken discourse

Joel Adams[1], **Steven Bedrick**[1], **Gerasimos Fergadiotis**[2], **Kyle Gorman**[3] and **Jan van Santen**[1]

[1]Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR
[2]Speech & Hearing Sciences Department, Portland State University, Portland, OR
[3]Google, Inc., New York, NY

Abstract

We present a system for automatically detecting and classifying phonologically anomalous productions in the speech of individuals with aphasia. Working from transcribed discourse samples, our system identifies neologisms, and uses a combination of string alignment and language models to produce a lattice of plausible words that the speaker may have intended to produce. We then score this lattice according to various features, and attempt to determine whether the anomalous production represented a phonemic error or a genuine neologism. This approach has the potential to be expanded to consider other types of paraphasic errors, and could be applied to a wide variety of screening and therapeutic applications.

1 Introduction

Aphasia is an acquired neurogenic language disorder in which an individual's ability to produce or comprehend language is compromised. It can be caused by a number of different underlying pathologies, but can generally be traced back to physical damage to the individual's brain: tissue damage following ischemic or hemorrhagic stroke, lesions caused by a traumatic brain injury or infection, etc. It can also be associated with various neurodegenerative diseases, as in the case of Primary Progressive Aphasia. According to the National Institute of Neurological Disorders and Stroke, approximately 1,000,000 people in the United States suffer from aphasia, and aphasia is a common consequence of strokes (prevalence estimates for aphasia among stroke patients vary, but are typically in the neighborhood of 30% (Engelter et al., 2006)).

Anomia is a the inability to access and retrieve words during language production, and is a common manifestation of aphasia (Goodglass and Wingfield, 1997). An anomic individual will experience difficulty producing words and naming items, which can cause substantial difficulties in day-to-day communication.

The process of screening for, diagnosing, and assessing anomia is typically manual in nature, and requires substantial time, labor, and expertise. Compared to other neuropsychological assessment instruments, aphasia-related assessments are particularly difficult to computerize, as they typically depend on subtle and complex linguistic judgments about the phonological and semantic similarity of words, and also require the examiner to interpret phonologically disordered speech. Furthermore, the most commonly used assessments focus for practical reasons on relatively constrained tasks such as picture naming, which may lack ecological validity (Mayer and Murray, 2003).

In this work, we describe an approach to automatically detecting and analyzing certain categories of word production errors characteristic of anomia in connected speech. Our approach is a first step towards an automated anomia assessment tool that could be used cost effectively in both clinical and research settings,[1] and could also be applied to other disorders of speech production. The method we propose uses statistical language models to identify possible errors, and employs a phonologically-informed edit distance model to determine phonological similarity between the subject's utterance and a set of plausible "intended words." We then apply machine learning techniques to determine which of several categories a given erroneous production may fall into. We

[1]As in the computer-administered (but manually-scored) assessments developed by Fergadiotis and colleagues (Fergadiotis et al., 2015; Hula et al., 2015).

Proceedings of the BioNLP 2017 workshop, pages 1–8,
Vancouver, Canada, August 4, 2017.

show results on intrinsic evaluations comparable to state-of-the-art sentence completion, as well as an extrinsic measure of classification well above a "most-frequent" baseline strategy.

1.1 Anomia and Paraphasias

Anomia can take several different forms, but in this work we are concerned with *paraphasias*, which are unintended errors in word production.[2]

There are several categories of paraphasic error. *Semantic errors* arise when an individual unintentionally produces a semantically-related word to their original, intended word (their "target word"). A classic semantic error would be saying "cat" when one intended to say "dog."

Phonemic (sometimes called "formal") errors occur when the speaker produces an unrelated word that is *phonemically related* to their target: "mat" for "cat", for example. It is also possible for an erroneous production to be *mixed*, that is both semantically and phonemically related to the target word: "rat" for "cat." Individuals with anomia also produce *unrelated* errors, which are words that are neither semantically or phonemically related to their intended target word: for example, producing "skis" instead of "zipper."

Each of these categories shares the commonality that the word produced by the individual is a "real" word. There is another family of anomic errors, *neologisms*, in which the individual produces *non-word* productions. A neologistic production may be phonemically related to the target, but containing phonological errors: "[dɑɪnoʊsɔɪ]" for "dinosaur." These are often referred to as *phonological* paraphasias. Alternatively, the individual may produce *abstruse neologisms*, in which the produced phonemes bear no discernable similarity to any "real" lexical item ("[æpməl]" for "comb"[3]).

The present work focuses exclusively on neologisms, both of the phonological variety as well as the abstruse variety. However, our fundamental approach can be extended to include other forms, as described in section 6.

Typical methods of diagnosing, staging, and otherwise characterizing anomia involve determining the number and kinds of paraphasias produced by an individual while undergoing some structured language elicitation process, for example a confrontation naming test (see (Kendall et al., 2013) and (Brookshire et al., 2014) for examples of such a study). As alluded to previously, producing these counts and classifications is a complex and laborious process. Furthermore, it is also often an inherently subjective process: are "carrot" and "banana" semantically related? What about "hose" and "rope"?

Reliability estimates of expert human performance at paraphasia classification in confrontation naming scenarios reflect the difficulty in this task. One recent study reported a kappa-equivalent score of 0.76 — a score that that is certainly acceptable, but that leaves much room for disagreement on the status of specific erroneous productions (Minkina et al., 2015). Other reported scores fall in a similar range (Kristensson et al., 2015), including when the productions are from neurotypical individuals (Nicholas et al., 1989). Automating this aspect of the task would not only improve efficiency, but would also decrease scoring variability. Having a reliable, automated method to analyze paraphasic errors would also expand the scope of what is currently possible in terms of assessment methodologies.

Notably, the approach we outline in this paper is explicitly designed to work on samples of natural, connected speech. It builds upon previous work by Fergadiotis et al. (2016) on automated analysis of errors produced in confrontation naming tests, and extends it into the realm of naturalistic discourse. It is our hope that, by enabling automated calculation of error frequencies and types on narrative speech, we might make using such material far easier in practice than it is today.

2 Data

For this work, we use the data set provided by the AphasiaBank project (MacWhinney et al., 2011), which has assembled a large database of transcribed interactions between examiners and people with aphasia, nearly all of whom have suffered a stroke. Notably, AphasiaBank also includes transcribed sessions with neurotypical controls. Each interaction follows a common protocol and script,

[2] Note that individuals *without* any sort of language disorder do occasionally produce errors in their speech; this fact has led to a truly shocking amount of study by linguists. Frisch & Wright (2002) provide a reasonable overview of the background and phonology of the phenomenon.

[3] This example was taken from a corpus of responses to a confrontation naming test (Mirman et al., 2010), in which the subject is shown a picture and required to name its contents. As such, in the case of this specific error, we have *a priori* knowledge of what the target word "should" have been. Obviously, in a more naturalistic task or setting, we would not have this advantage.

and is transcribed in great detail using a standardized set of annotation guidelines. The transcripts include word-level error codes, according to a detailed taxonomy of errors and associated annotations. In the case of semantic, formal, and phonemic errors, the word-level annotations include a "best guess" on the part of the transcriber as to what the speaker's intended production may have been.

Each transcribed session consists of a prescribed sequence of language elicitation activities, including a set of personal narratives (e.g.,"Do you remember when you had your stroke? Please tell me about it."), standardized picture description tasks, a story retelling task (involving the story of *Cinderella*), and a procedural discourse task.

We noted that the distribution of errors within sentences seems to obey the power law , with the majority of error-containing sentences containing a single error, followed somewhat distantly by sentences containing two errors, with a relatively steep dropoff thereafter. For the present study, we restricted our analysis to sentences that contained a single error. Our reasoning for this restriction was that we do not presently have a theoretically-informed model of what, if any, relationship there may be between multiple errors within a sentence. However, it seems quite likely that the errors occurring in a sentence containing (for instance) five paraphasic errors might be somehow related to one another. We anticipate exploring this phenomenon in the future (see section 6).

We chose to restrict our data to the story retelling task due to the constrained and focused vocabulary of the Cinderella story. This resulted in $\approx$ 1000 sentences from 385 individuals. We then identified sentences containing instances of our errors of interest: phonological paraphasia (AphasiaBank codes "p:n", "p:m") or abstruse neologism ("n:uk" and "n:k").

3 Methods

We first tokenized the AphasiaBank data using a modified version of the Penn Treebank tokenizer which left contractions as a single word and simply removed the connecting apostrophe, as these occasionally appear as target words and thus we needed to treat them as a single token. We left stopwords intact, and case-folded all sentences to upper-case. Cardinal numbers were collapsed into a category token, as were ordinal numbers and dates (each category was given its own token). The AphasiaBank transcripts include detailed IPA-encoded representations of neologistic productions, but any "real-world" usage scenario for our algorithm is unlikely to benefit from such high-quality transcription. We therefore translated the non-lexical productions into a simulated "best-guess" orthographic representation of the subject's non-lexical productions.

We next turned to the question of identifying neologisms in our sentences. Simply using a standard dictionary to determine lexicality could result in numerous "false positives," driven largely by proper names of people, brands, etc. To avoid this, we used the SUBTLEX-US corpus (Brysbaert and New, 2009) to identify neologisms. SUBTLEX-US was build using subtitles from English-language television shows and movies, and Brysbaert and New have demonstrated that it correlates with a number of psycholinguistic behavior measures (most notably, naming latencies) better than better-known frequency norms such as those derived from the Brown corpus or CELEX-2.

Upon identifying a possible non-word production, recall that our next goal is to determine whether it represents a *phonemic* error (substituting "[dɑɪnoʊsɔɹ]" for "dinosaur") or an *abstruse neologism* (a completely novel sequence of phonemes that does not correspond to an actual word). To help accomplish this, we train a language model to identify plausible words that *could* fit in the slot occupied by the erroneous production, and produce a lattice of these candidate target words (i.e., words that the subject may have been intending to produce, given what we know about the context in which they were speaking).

Our language models for this study were built using the New York Times section of the Gigaword newswire corpus (Parker et al., 2011). After success in preliminary experiments, we filtered this corpus by first training a Latent Dirichlet Allocation (LDA) topic model on the corpus using Gensim (Řehůřek and Sojka, 2010) over 20 topics. We then projected the text of each of the Cinderella narrative samples into the resulting topic space, and calculated the centroids for the narrative task. This yielded a subset of the larger corpus of New York Times articles that was "most similar" to the Cinderella retellings, and we used these to build our language models.

We investigated two different language model-

ing approaches: a traditional FST-encoded ngram language model, and a neural-net based log-bilinear (LBL) language model. For the FST representation, we used the the OpenGrm-NGram language modeling toolkit (Roark et al., 2012) and used an n-gram order of 4, with Kneser-Ney smoothing (Kneser and Ney, 1995). For the LBL approach, we used a Python implementation[4] of the language model described by Mnih and Teh (Mnih and Teh, 2012). We used word embeddings of dimension 100, and a 5-gram context window. In both cases we trained two language models: one trained on the "task-specific" subset of Gigaword, and another trained on the AphasiaBank control data. We combined these with a simple mixing coefficient, λ as shown in Equation 1 where $P_{GW}(w)$ is the language model probability of word w computed against the Gigaword corpus and the $P_{AB}(w)$ is the language model probability trained on the AphasiaBank controls.

$$P(w) = \lambda \cdot P_{AB}(w) + (1 - \lambda) \cdot P_{GW}(w) \quad (1)$$

We evaluate non-lexical productions as follows. First, we use the Phonetisaurus grapheme-to-phoneme toolkit (Novak et al., 2012) to translate our orthographic representation into an estimated phoneme sequence. We then calculate a phonologically-aware edit distance between each non-word production and every word in our lexicon up to some maximum edit distance (in our case 4.0). Phonemes from a related class (e.g. vowels) are considered lower cost replacements than those from another class (e.g. unvoiced fricatives). This gives us a set of candidates which are phonologically similar to the production.

We next used our language models to produce lattices representing a set of possible sentences that the subject could plausibly have been intending to produce. At the point in the produced sentence where our error detection system indicated that a non-word production occurred, we represent the anomaly by the union of all possible words in our edit-distance constrained lexicon (see figure 3 for an example sentence lattice). Finally, we use the language models to score the resulting sentence lattice so as to be able to rank the candidate words, and use the estimated sentence-level probability for each candidate word (i.e., the hypothesized intended utterance featuring that word). Put simply,

[4] https://github.com/ddahlmeier/neural_lm

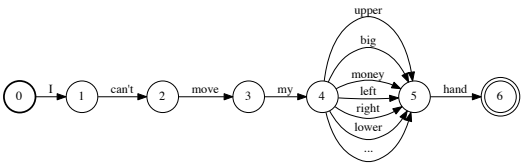

Figure 1: An example candidate word lattice for the production "I can't move my [vɑɪ] hand."

for each candidate intended word, we produce a version of the subject's utterance with that hypothesized word in place of the anomalous utterance, and score this hypothesized utterance with the language model.

At this point in the process, we have the following information about each erroneous production: a best-guess orthographic transcription of what the individual actually produced, and a ranked list of plausible words that they could potentially have been attempting to produce, together with probability estimates for each hypothesized production.

To determine the category of our error productions— again, between productions representing phonological errors such as "[dɑɪnoʊsɔɹ]" for "dinosaur", and productions representing abstruse neologisms— we trained a binary classifier using features representing the characteristics of the candidate word space surrounding the erroneous production. Our intuition is that phonemic errors were much more likely than abstruse neologisms to have highly-ranked candidate target words that were *also* phonologically similar to the subject's actual production.

To capture this, we performed the following procedure for each error-containing utterance. We first divide our list of candidate intended words into buckets by edit distance (0.5, 1.0, 1.5, etc.[5]). Each bucket can now be thought of as a ranked list of probabilities, each representing a possible hypothesized intended utterance featuring a word within that bucket's edit distance of the actual (anomalous) utterance.

We next represent each bucket with a feature vector consisting of the count of words in that

[5] Recall that our phonological edit distance metric allows for partial costs for related phonological substitutions.

bucket, as well as descriptive statistics regarding the distribution of language model probabilities in that bucket (min, max, etc.). We then concatenate each bucket's features together into a master feature vector for the utterance. Our expectation is that these features will be relatively evenly distributed across buckets in the case of utterances containing abstruse neologisms, whereas utterances featuring phonological paraphasias will vary according to phonological edit distance.

Once we have computed feature vectors for each utterance, we used the Scikit-learn Python machine learning library (Pedregosa et al., 2011) to train a Support Vector Machine classifier to distinguish between utterances phonological and abstruse neologisms. We evaluate its performance using leave-one-out cross-validation.

4 Results

We perform two evaluations of our model: an intrinsic evaluation of how often our system includes the target word in the top-n ranked candidates, and an extrinsic evaluation where we attempt to classify a paraphasia between phonological errors and abstruse neologisms.

Our motivation for evaluating our system's performance on target word prediction is tied to our classification assumptions. In an ideal case for a phonological error, the target word should fall within one of the buckets representing a low edit distance. If our language model successfully rates the target as likely, we would see an high maximum probability within that bucket, which is a feature in our classifier.

The performance of our language models on the top-n ranked evaluation can be seen in Table 1. The log-bilinear model outperformed the FST in all cases. This finding is similar to state of the art results for automatic sentence completion systems–particularly for phonemic errors–as we'll discuss in greater detail in Section 5. Both systems did a better job of predicting the target word for phonemic errors than they did for abstruse neologisms. It's not immediately clear what the reason for this is. However, anecdotally, sentences including abstruse neologisms are also often agrammatical.

For the evaluation of our classification, we created a simple majority class baseline classifier that always chooses the largest class of errors (phonemic errors in this case). This baseline classifier has a classification accuracy of .51. The results of classification can be seen in Table 2. Both of our systems handily outperformed baseline: the FST by a relative 20% improvement, and the LBL nearly 33%. As we expected from the top-n results, classification based on the LBL outperformed that based on the FST.

Error	n	FST	LBL
Phonemic	1	.43	.52
Phonemic	5	.54	.66
Phonemic	10	.59	.69
Phonemic	20	.67	.77
Phonemic	50	.72	.81
Abstruse Neo.	1	.29	.35
Abstruse Neo.	5	.41	.49
Abstruse Neo.	10	.44	.51
Abstruse Neo.	20	.51	.59
Abstruse Neo.	50	.54	.60

Table 1: Accuracy of language model predicting the correct target word within the first n results.

Features	FST	LBL
count, mean	.612	.661
count, mean, max	.621	**.680**
count, mean, max, min	.610	.659
count, mean, max, dist.	.610	.659

Table 2: Classification accuracy by model. Baseline accuracy of choosing the most common error type is .510.

The "dist" feature listed in table 2 is the edit distance of a given bucket normalized by the number of phonemes in the actual error production. It was not found to be productive as a feature, nor was the minimum language model probability of words in a given bucket ("min" in the table). The best results for both systems were a combination of count of candidate terms per bucket ("count") concatenated with the maximum and mean language model probabilities within a bucket ("max" and "min", respectively).

We varied the mixing-coefficient (λ) from Equation 1 in both the FST and LBL approaches. As long as the resulting model includes a non-trivial weighting of the Cinderella corpus (typically anything better than $\lambda = 3$), the actual value of the mixing coefficient was irrelevant to either of our evaluations. In this, it appears to work as designed, with the Gigaword corpus providing background probabilities, and the AphasiaBank Cinderella con-

trol retellings increasing the weight of topically important words that are otherwise rare (like "Cinderella" and "carriage").

5 Related Work & Discussion

As far back as Shannon's word-guessing game (Shannon, 1951), researchers have sought to leverage the statistical regularities in natural language to predict missing or subsequent words. In practice, however, this proves to be a surprisingly challenging problem. Language occurs at levels beyond simply choosing lexical items, and local statistical characteristics of language often fail to capture syntactic and semantic patterns. Zweig & Burges (2012) provide an enlightening discussion on the limitations of relying on n-gram guessing for syntactically complex tasks such as "identify the missing word in the sentence," and also describe a very challenging language model evaluation task built around this problem. They tested a variety of language modeling approaches using their task, and report that well-trained generative n-gram models achieve correct predictions $\approx$ 30% of the time. State-of-the-art performance on the their word prediction task using recurrent neural network langage models,[6] report highest scores are in the mid-50% range (Mirowski and Vlachos, 2015; Mnih and Kavukcuoglu, 2013).

In our case, the nature of our data renders this task even more challenging. Our sentences are often short and agrammatical (often missing or misusing determiners, for example), and are produced by individuals with impaired language ability.

As such, our results are actually quite similar to those reported in recent literature. Our average accuracy of our FST n-gram (over both classes of errors) selecting the appropriate word is $\approx$ 35% while our LBL model's performance of $\approx$ 43% is comparable to the 5-gram LBL performance of 49.3 reported by Mnih and Teh on the MSR Sentence Completion Challenge dataset (Mnih and Teh, 2012).

6 Conclusion & Future Work

While the system's performance is quite good on both intrinsic and extrinsic evaluation, there remains much interesting work left to do on the problem.

We currently only evaluate sentences with a single error, and more generally have not investigated whether sentences with multiple errors are different from mono-error sentences in terms of error distribution or structure. However, our approach *should* be able to generalize to sentences with additional errors, and we will be investigating this in future work.

Additionally, the AphasiaBank transcripts include phrasal dependency and part-of-speech tags which we are currently not using. In future work we will investigate including these as features in language modelling, as there is some evidence that this improves the conceptually related task of contextual spellcheck(Fossati and Di Eugenio, 2008).

There is quite a bit of work that can be done on the language models as well. A more nuanced approach to topic adaptation is worth investigating, and we plan to experiment with using non-newswire corpora. Despite our attempts to focus the corpus via LDA, there is a major difference between the written language of the New York Times, and the spoken dialogue between the aphasic subjects and their clinicians.

The most exciting area for further research is the inclusion of semantic information in our classification. While our topic-specific language model provides our model with some implicit semantic information, a more principled approach to semantic relevance could potentially improve the classification of phonemic errors versus abstruse neologisms by determining whether a given candidate word is semantically relevant in context. More intriguingly, it would give us a way to start investigating semantic errors, and those errors that include "real" words (for example, the previously discussed error of replacing "dog" with "cat").

One major limitation of our current system is its reliance on human-produced transcriptions of speech samples. In practice, transcription is rarely feasible in clinical settings, and even in research settings is often challenging, which may seem to limit the applicability of our approach. Notably, however, our system does not require detailed *phonetic* transcription, and merely requires "best-guess" orthographic transcription of neologisms. As such, one could in principle use automatic speech recognition (ASR) to analyze a recording of a patient or research subject, and produce a transcript on which our methods could be

[6] See De Mulder et al. (2015) for a recent review on this subject.

run.[7] Fraser et al. (2015) have had some success at applying ASR to samples of aphasic speech and performing downstream analysis on the resulting transcripts, and we anticipate experimenting with similar techniques in the future.

Acknowledgments

We thank the BioNLP reviewers for their helpful comments and advice. This material is based upon work supported in part by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under awards R01DC012033 and R03DC014556. The content is solely the responsibility of the authors and does not necessarily represent the official views of the granting agencies or any other individual.

References

C. E. Brookshire, T. Conway, R. H. Pompon, M. Oelke, and D. L. Kendall. 2014. Effects of intensive phonomotor treatment on reading in eight individuals with aphasia and phonological alexia. *American Journal of Speech-Language Pathology* 23(2):S300–S311.

M. Brysbaert and B. New. 2009. 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.

Wim De Mulder, Steven Bethard, and Marie-Francine Moens. 2015. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language* 30(1):61–98.

S. T. Engelter, M. Gostynski, S. Papa, M. Frei, C. Born, V. Ajdacic-Gross, F. Gutzwiller, and P. A. Lyrer. 2006. Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis. *Stroke* 37(6):1379–1384.

G. Fergadiotis, S. Kellough, and W. D. Hula. 2015. Item Response Theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research* 58(3):865–877.

Gerasimos Fergadiotis, Kyle Gorman, and Steven Bedrick. 2016. Algorithmic Classification of Five Characteristic Types of Paraphasias. *American Journal of Speech-Language Pathology* 25(4S):S776–12.

Davide Fossati and Barbara Di Eugenio. 2008. I saw tree trees in the park: How to correct real-word spelling mistakes. In *LREC*.

[7] Depending on the specifics of the ASR system, it may in fact be possible to retain phonological information, which, while not necessary, certainly could be helpful to our system.

K. C. Fraser, N. Ben-David, G. Hirst, N. Graham, and E. Rochon. 2015. Sentence segmentation of aphasic speech. In *ACL*. pages 862–871.

S. A. Frisch and R. Wright. 2002. The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics* 30(2):139–162.

H. Goodglass and A. Wingfield. 1997. *Anomia: Neuroanatomical and cognitive correlates*. Academic Press, New York.

W. D. Hula, S. Kellough, and G. Fergadiotis. 2015. Development and simulation testing of a computerized adaptive version of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research* 58(3):878–890.

D. L. Kendall, R. H. Pompon, C. E. Brookshire, I. Minkina, and L. Bislick. 2013. An analysis of aphasic naming errors as an indicator of improved linguistic processing following phonomotor treatment. *American Journal of Speech-Language Pathology* 22(2):S240–S249.

R Kneser and H Ney. 1995. Improved backing-off for M-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, pages 181–184.

J. Kristensson, I. Behrns, and C. Saldert. 2015. Effects on communication from intensive treatment with semantic feature analysis in aphasia. *Aphasiology* 29(4):466–487.

B. MacWhinney, D. Fromm, M. Forbes, and A. Holland. 2011. AphasiaBank: Methods for studying discourse. *Aphasiology* 25(11):1286–1307.

J. Mayer and L. Murray. 2003. Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology* 17(5):481–497.

I. Minkina, M. Oelke, L. P. Bislick, C. E. Brookshire, R. Hunting Pompon, J. P. Silkes, and D. L. Kendall. 2015. An investigation of aphasic naming error evolution following phonomotor treatment. *Aphasiology* epub ahead of print.

D. Mirman, T. J. Strauss, A. Brecher, G. M. Walker, P. Sobel, G. S. Dell, and M. F. Schwartz. 2010. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology* 27(6):495–504.

Piotr Mirowski and Andreas Vlachos. 2015. Dependency Recurrent Neural Language Models for Sentence Completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 511–517.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 2265–2273.

Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426* .

L. E. Nicholas, R. H.and Maclennan D. L. Brookshire, J. G. Schumacher, and S. A. Porrazzo. 1989. Revised administration and scoring procedures for the Boston Naming Test and norms for non-brain-damaged adults. *Aphasiology* 3(6):569–580.

J. R. Novak, N. Minematsu, and K. Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *International Workshop on Finite State Methods and Natural Language Processing*. pages 45–49.

R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda. 2011. English Gigaword 5th Edition. Linguistic Data Consortium: LDC2011T07.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, and T. Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *ACL*. pages 61–66.

C. Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 50:50–64.

Geoffrey Zweig and Chris J C Burges. 2012. A Challenge Set for Advancing Language Modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, Montréal, Canada, pages 29–36.

R. Řehůřek and P. Sojka. 2010. Software framework for topic modelling with large corpora. In *LREC*. pages 45–50.