UtahPOET: Disorder mention identification and context slot filling with cognitive inspiration

Kristina Doing-Harris
Department of Biomedical Informatics
University of Utah Health Sciences Center
421 Wakara Way, Ste 140
Salt Lake City, UT 84108 USA
kristina.doing-harris@utah.edu

Sean Igo
Department of Biomedical Informatics
University of Utah Health Sciences Center
421 Wakara Way, Ste 140
Salt Lake City, UT 84108 USA
Sean.igo@utah.edu

Jianlin Shi
Department of Biomedical Informatics
University of Utah Health Sciences Center
421 Wakara Way, Ste 140
Salt Lake City, UT 84108 USA
Jianlin.shi@utah.edu

John Hurdle
Department of Biomedical Informatics
University of Utah Health Sciences Center
421 Wakara Way, Ste 140
Salt Lake City, UT 84108 USA
John.hurdel@utah.edu

Abstract

We describe the performance of UtahPOET on SemEval 2015 Task 14. UtahPOET is a cognitively inspired system designed to extract semantic content from general clinical texts. We find that our system performs much better on the context slot-filling aspects of Tasks 2A and 2B than the disorder CUI mapping of Tasks 1 and 2B or the body location CUI mapping of Task 2B. Our problems with CUI mapping suggested several possible system improvements. An alteration in the correspondence between the system architecture and psycholinguistic findings is also indicated.

1 Introduction

We note at the outset that our team approaches clinical NLP using a new, cognitively inspired architecture. We value dataset independence, so our design priorities do not completely overlap those encompassed by the goals of Task 14. We share the SemEval vision of extracting the full semantic content of clinical text. Our short-term goal, however, was to field test an early prototype of our new architecture and Task 14 provided a convenient and well-designed use case.

1.1 Cognitive inspirations

Only the human brain is currently able to extract full semantic content from text. We propose an intermediate step between artificial neurons (Merolla et al., 2014; Sowa, 2010) and statistical machine learning (ML). We use ML and rule-based NLP components with demonstrated success in clinical information extraction arranged in an architecture inspired by well-documented findings with respect to cortical processing.

Briefly, UtahPOET is inspired by findings related to: layered cognitive processes, the distinction between the dorsal and ventral language processing streams, and the phenomenon of iterative refinement. The type of layered (i.e., staged or hierarchical) processing we use shares much in common with traditional NLP and biologically inspired cognitive architectures (Chella, Cossentino, Gaglio, & Seidita, 2012; Indurkhya & Damerau, 2010; Sowa, 2010). We will discuss our system's layering in the system description below.

Our distinctive model of dorsal-ventral processing streams comes from psycholinguistic findings. The interpretation of unfamiliar or ungrammatical constructions, rule-based processing, and learning have been linked to dorsal processing streams in the brain. Ventral processing streams handle familiar, expected, regular con-

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 399–405,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

structions as well as heuristic-type processing (Dominey & Inui, 2009; Hickok & Poeppel, 2004; Kellmeyer et al., 2013; Levy et al., 2009; Price, 2013; Yeatman, Rauschecker, & Wandell, 2013). Iterative refinement is the repeated application of top-down processing during bottom-up processing. In Cognitive Science top-down and bottom-up refer, in essence, to processes that rely on previous knowledge and those that do not, respectively (Traxler, 2012).

Top-down processing is evident in each stage of an NLP pipeline, e.g., "knowing" how the end of a sentence is marked. We see combining world knowledge with the outcome of one processing stage and then using that to update the outcome of a previous stage as iterative refinement. This resembles how humans 're-parse' garden path sentences (McKoon & Ratcliff, 2007).

The UtahPOET approaches solving semantic extraction problems by enabling dependency parsing. However, ungrammatical text is common in clinical notes (Fan et al., 2013; Meystre, Savova, Kipper-Schuler, & Hurdle, 2008). This text often "breaks" dependency parsers, so we process grammatical and ungrammatical text separately. Dependency parsing is useful because it exploits world knowledge about the structure of English sentences. As such, it simplifies the processing of conjunctions and the aggregation of words and relationships, particularly those separated in the text, without supervised training. Retaining sentence structure allows dataset independence and latitude in future relationship finding.

1.2 Considerations for evaluation

We propose a couple of considerations useful for evaluating NLP systems' results under Task 14. The current evaluation includes strict matching to a Gold Standard set of Unified Medical Library System (UMLS) Metathesaurus (Browne, Divita, Aronson, & McCray, 2003) CUIs. We think this standard leads to over-fitting the data, which leads to less generally useful systems. Clinical terms do not guarantee a one-to-one correspondence between term and referent. A point demonstrated by inter-annotator agreement of anything less than 100%.

The redundancy of the UMLS Methathesaurus further undermines strict CUI mapping. Redundancy is best illustrated by body location mapping. Within the UMLS semantic types relevant to body location are T023 (Body part, organ or organ component) and T029 (Body location or region). We notice inconsistency in the Gold Standard in the use of these semantic types. For one document annotators chose 'Pericardial sac structure (T023)' over "Pericardial body location (T029)', while in another annotators preferred 'Neck (T029)' over 'Entire neck (T023).'

Partial matches create problems as well. The Task evaluation only considers partial span matches correct if the CUI for the full match is reported. However, if the span is only partially matched the correct CUI should change. For example, the mapping 'Left ventricular hypertrophy' to C0149721, when partially matched with 'Ventricular hypertrophy' would seem to be more correctly mapped to C0340279.

2 System description

The UtahPOET system is built in Apache UIMA (Ferrucci & Lally, 1999). It has the layered structure common to NLP pipelines (see Figure 1). The pre-processing stage finds sentence boundaries (stages A), breaks the sentence into tokens (stage B), and assigns each token a part-of-speech (POS) tag (stage B).

2.1 Dorsal-ventral stream separation and iterative refinement

After preprocessing, we add stages to begin dorsal and ventral separation and iterative refinement. In stage C, we divide dorsal and ventral streams by separating ungrammatical and grammatical text. We refer to ungrammatical text as *nonprose qs_segments*. Nonprose is differentiated from prose (well-formed sentences) by two rules. First, well-formed sentences contain at least one verb. Second, well-formed sentences do not contain more than four numbers (e.g., labs) per verb.

Iterative refinement occurs in Stage D. Realizing that standard sentence segmentation may not perform well with nonprose (e.g., consider common lists like medications with no periods), we then re-segment the text breaking each nonprose qs_segment at the next carriage return, line break, or end-of-line character. The dotted line in Figure 1 signifies that it is a repeated process.

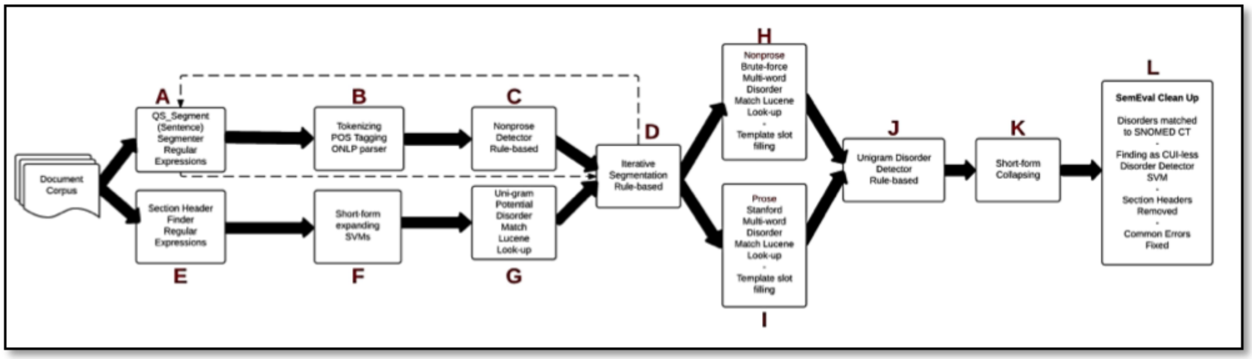

A
QS_Segment (Sentence) Segmenter Regular Expressions
B
Tokenizing POS Tagging ONLP parser
C
Nonprose Detector Rule-based
D
Iterative Segmentation Rule-based
Document Corpus
Section Header Finder Regular Expressions
E
Short-form expanding SVMs
F
Uni-gram Potential Disorder Match Lucene Look-up
G
H
Nonprose Brute-force Multi-word Disorder Match Lucene Look-up - Template slot filling
I
Prose Stanford Multi-word Disorder Match Lucene Look-up - Template slot filling
J
Unigram Disorder Detector Rule-based
K
Short-form Collapsing
L
SemEval Clean Up
Disorders matched to SNOMED CT - Finding as CUI-less Disorder Detector SVM - Section Headers Removed - Common Errors Fixed

Figure 1. The overall UIMA pipeline for UtahPOET (please zoom for readability).

2.2 UtahPOET specific parallel ‘preprocessing’

UtahPOET has section header identification and short-form expansion processes that run parallel to the ‘pre-processing’ stages. These stages are E and F in Figure 1.

In stage E regular expressions are used to identify section headers. The regular expression rules are found using automatic regular expression extraction (Bui & Zeng-Treitler, 2014).

In stage F, a series of SVMs are used to expand short forms. The feature vectors for these SVMs include context vectors as bags-of-words and section headers. The short form-long form pairs are extracted from the ADAM dataset (Zhou, Torvik, & Smalheiser, 2006) but limited to clinical terms. One classifier is trained for each ambiguous normalized short form that has multiple corresponding long forms. Classifiers are trained using the UMN clinical abbreviation and acronym sense inventory (Moon, Pahkhomov, Liu, Ryan, & Melton, 2014) and context information retrieved from PubMed case reports. The features are built on LVG (Browne et al., 2003) normalized bag of word, section header and short form string. The expanded short forms are inserted into the original text, preserving the original span information in UIMA annotations for span matching back to original text in the final stage.

2.3 Disorder detection in dorsal and ventral streams

Stage G has two purposes: to identify single-word disorder terms and to limit the number of words that will be looked up in later stages. After stopwords are removed, each word in the document is stemmed using LVG (Browne et al., 2003) and fetched from a Lucene index made from the UMLS Metathesaurus restricted to the clinical sources indicated in (Wu et al., 2012), including SNOMEDCT, MSH, NCI, RDC, MTH, SNMI, MDR, SCTSPA, CHV, CCPS. The sematic types included reflect disorders, body locations, and modifiers. Modifiers include qualitative, quantitative and spatial concepts.

For the identification of multi-word terms and context slot filling in stages H and I, we split the text segments based on the previously described nonprose (stage H) prose (stage I) distinction. The dorsal stream is associated with rule-based processing. In this case the rule associated with nonprose qs_segments, is that adjacent unigram disorder terms are likely to be part of a multi-word term. Equivalently, the body location and severity relevant to a disorder will be adjacent to the disorder mention. The ventral processing stream exploits world knowledge about regularity of construction by dependency parsing. Unigram matches that share dependencies are likely to be part of a multi-word term and reflect relevant body locations and severities.

In both stages (H and I), we build as long a multi-word term as possible then attempt to match the term to a Lucene index into the UMLS Metathesaurus restricted to the clinical sources listed above and only the disorder semantic types. If the term does not match, it is incrementally reduced token-by-token, with all combinations of words checked for a match at each step.

Context slots are filled by overwriting entries in a default template: the mention is not negated, the subject is the patient, the mention is not uncertain, severity and course are unmarked, the mention is not conditional or generic, and there is no body location given.

Negation, uncertainty, subject, and generic mention are found at the sentence level in nonprose and the dependency level in prose by looking for specific text. The remaining slot values were located by adjacency (nonprose) or dependency (prose).

2.4 Post-processing

Stage K takes place outside of UIMA. It collapses expanded short-forms back to their original spans and updates spans of all the other annotations in the file so our output spans reflect those from the SemEval gold standard. Stage L (SemEval clean up) is the final stage of the pipeline in Figure 1. Here we map, where possible, disorder CUIs from SNOMED CT. This stage also incorporates a process for identifying terms matched to the UMLS Metathesaurus semantic type finding (T033) that are considered CUI-less disorders in the SemEval gold standard. We use a structured SVM to classify the spans of *findings* to CUI-less disorder or not. We used the Cornell SVM^{struct} SVM^{hmm} model. (Joachims, n.d.) Feature vectors are 4-word context-window (2 before and 2 after), bag-of-words stemmed with stopwords removed using NLTK (Bird, Loper, & Klein, 2009). The SVM parameters were slack vs. weight vector magnitude (-c) of 25000 and epsilon (-e) of 0.5.

This stage also removes all disorders found within section headers as well as annotations that reflect either spurious UMLS Metathesaurus mappings or problems with short-form expansion.

3 Results

UtahPOET was not expected to perform well on either Task 1 or Task 2A. In both cases, our unwillingness to adhere to the gold standard CUIs caused us to score at the bottom of the pack. Sixteen teams competed in Task 1. We were 15th. Only 6 teams competed in Task 2A, we were last. Considering the context slot filling, apart from CUI and body location, in Task 2A would have moved us up one rank.

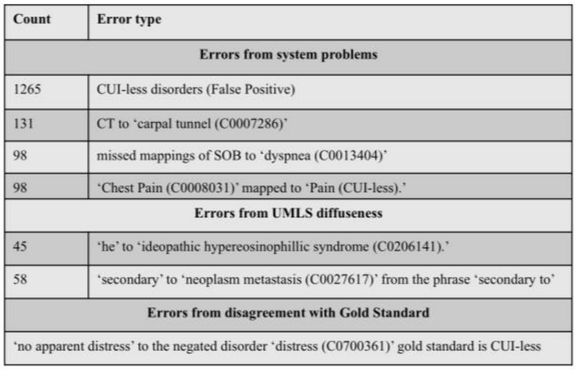
We were mainly focused on Task 2B where we scored in the middle of the pack until many of the teams withdrew. Nine teams remain in the Task 2B competition. Our three runs come second to the last. Again looking at only slot filling, we would have moved up three ranks.

Our results for the development set closely mirrored those on the test set; so will not be described.

3.1 Difference between runs

We were unsure whether scoring favored F-scores or accuracy so we submitted runs favoring one or the other. For both tasks, we submitted 2 copies of our best run in case there was a problem creating one of the submissions. If one failed, there would still be one left. In tasks 1 and 2A runs 1 and 2 were the same. Run 3 had a stricter Lucene match leading to higher accuracy and lower F-score (i.e., reduced numbers of true positive, false positive and false negative concepts). The stricter match required that only the words found in the document appear in the matched term, no extra words were allowed. Thus, “hypertension” would not match the UMLS Metathesaurus entry “hypertensive disease.” In task 2B, runs 2 and 3 are the same. This time run 1 has a slightly higher accuracy, but lower F-score due to change in Lucene matching.

For task 2A, we also realized that we could use the gold standard spans to match the context found by UtahPOET without finding an associated concept, if we reported the span as a CUI-less disorder.



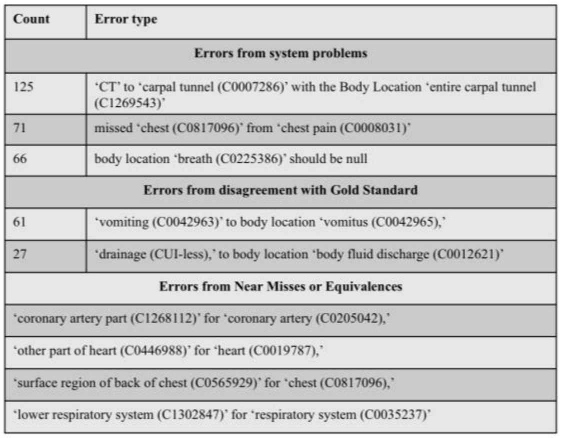
Count	Error type
Errors from system problems	
1265	CUI-less disorders (False Positive)
131	CT to ‘carpal tunnel (C0007286)’
98	missed mappings of SOB to ‘dyspnea (C0013404)’
98	‘Chest Pain (C0008031)’ mapped to ‘Pain (CUI-less).’
Errors from UMLS diffuseness	
45	‘he’ to ‘ideopathic hypereosinophillic syndrome (C0206141).’
58	‘secondary’ to ‘neoplasm metastasis (C0027617)’ from the phrase ‘secondary to’
Errors from disagreement with Gold Standard	
‘no apparent distress’ to the negated disorder ‘distress (C0700361)’ gold standard is CUI-less	

Table 2. Examples of CUI mapping error for disorders (please zoom for readability).

3.2 CUI and body location error analysis

Tables 2 and 3 list examples of the CUI mapping errors made by UtahPOET. For disorders, they fall into three increasingly large groups, system problems, UMLS diffuseness, and disagreement with the gold standard.

CUI-mapping errors in body location assignment were, in increasing order of size, due to system problems, disagreement with the gold standard and near misses or equivalences.



Count	Error type
Errors from system problems	
125	'CT' to 'carpal tunnel (C0007286)' with the Body Location 'entire carpal tunnel (C1269543)'
71	missed 'chest (C0817096)' from 'chest pain (C0008031)'
66	body location 'breath (C0225386)' should be null
Errors from disagreement with Gold Standard	
61	'vomiting (C0042963)' to body location 'vomitus (C0042965),'
27	'drainage (CUI-less),' to body location 'body fluid discharge (C0012621)'
Errors from Near Misses or Equivalences	
'coronary artery part (C1268112)' for 'coronary artery (C0205042),'	
'other part of heart (C0446988)' for 'heart (C0019787),'	
'surface region of back of chest (C0565929)' for 'chest (C0817096),'	
'lower respiratory system (C1302847)' for 'respiratory system (C0035237)'	

Table 3. Examples of CUI mapping error for body locations.

4 Discussion

The UtahPOET system can successfully extract semantic information from clinical text. The system construction has slightly different priorities than the Task organizers. Our priority of creating a dataset agnostic solution for semantic extraction problems prompted us to offer considerations for the evaluation and to look to cognitive findings for system design inspiration.

4.1 Implications for system improvement

Necessary system alterations are revealed by disorder CUI mapping error analysis in Table 3. CUI-less disorders are the most error prone. We will be adding features to the CUI-less disorder SVM to improve performance. Two mapping mistakes 'CT' and 'he' that may be fixed by a walk back to the most common form. We will investigate a method to implement a walk back. Standardizing the expanded long-forms would catch the missed 'SOB' mappings. Checking for phrase 'secondary to' would also be helpful.

We find support for our evaluation considerations above in CUI and body location mappings, which disagree with the gold standard. For example, if 'shortness of breath' is given the body location 'breath,' giving 'vomiting' to body location 'vomitus' and 'drainage' to location 'body fluid discharge' should be acceptable.

UtahPOET is prone to near misses. We see these near misses as a type of graceful degradation, which is a hallmark of cognitive systems. Graceful degradation is the ability to function despite making errors. Ferreira and Patson call this "good enough" processing (Ferreira & Patson, 2007).

4.2 Implications for cognitive architecture

The hierarchical layers from psycholinguistics are lexical, syntactic and semantic processing, which proceed in that order. We do not adhere strictly to this hierarchy. Many cognitive scientists think a proper hierarchy is unlikely (Frank, Bod, & Christiansen, 2012).

We were inspired to separate prose and nonprose based on the ventral-dorsal distinction between grammatical and ungrammatical text. It is tempting to equate heuristics with ML and rules with specific if…then statements. The cognitive science literature indicates that this is a mistake (Hahn & Chater, 1998). All heuristics are thought to start as rule-based. The rule-based decision is overlearned to the point of automaticity and called a heuristic. Therefore we do not use ML components in only one path.

Currently, UtahPOET leverages iterative refinement for sentence segmentation only. Once we implement greater integration with long-term memory (LTM) representation, we will have the facility to recognize clashes and implement more extensive iterative refinement. With our ML components, we can clearly see how learning requires its own pathway. Each of these systems is trained outside the UtahPOET pipeline and would require retraining, if new information were introduced.

Acknowledgments

We are grateful for the support of the National Library of Medicine grant R01LM010981. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. We would like to thank Duy Duc An Bui for building the Section Header component and Sarathkrishna Swaminathan for building the CUI-less disorder structure SVM.

References

Bird, Steven, Loper, Edward, & Klein, Edward. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Browne, Allen C., Divita, Guy, Aronson, Alan R., & McCray, Alexa T. (2003). UMLS Language and Vocabulary Tools: AMIA 2003 Open Source Expo, *2003*, 798.

Bui, Duy D. A., & Zeng-Treitler, Qing. (2014). Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association*, *21*(5), 850–857.

Chella, Antonio, Cossentino, Massimo, Gaglio, Salvatore, & Seidita, Valeria. (2012). A general theoretical framework for designing cognitive architectures: Hybrid and meta-level architectures for BICA. *Biologically Inspired Cognitive Architectures*, *2*(C), 100–108.

Doing-Harris, K.ristina Patterson, Olga, Igo, Sean, & Hurdle, John. (2013). Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts. Proceedings of the 7th international workshop on Data and text mining in biomedical informatics, ACM, 9-12.

Dominey, Peter F., & Inui, Toshio. (2009). Cortico-striatal function in sentence comprehension: Insights from neurophysiology and modeling. *Cortex*, *45*(8), 1012–1018.

Fan, Jung-wei, Yang, Elly W., Jiang, Min, Prasad, Rashmi, Loomis, Richard M., Zisook, Daniel S., et al. (2013). Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *Journal of the American Medical Informatics Association*, *20*(6), 1168-1177.

Ferreira, Fernanda, & Patson, Nikole D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Ferrucci, David, & Lally, Adam. (1999). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4), 327–348.

Frank, Stefan L., Bod, Rens, & Christiansen, Morten H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522-4531.

Hahn, Ulrike, & Chater, Nick. (1998). Similarity and rules: distinct? Exhaustive? Empirically distinguishable? *Cognition*, *65*(2-3), 197–230.

Hickok, Gregory, & Poeppel, David. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67–99.

Indurkhya, Nitin, & Damerau, Fred J. (Eds.). (2010). *Handbook of Natural Language Processing* (Second.). Chapman and Hall. p. 168.

Joachims, Thorston. (Ed.). *Cornell SVM^{struct}*. Retrieved January 30, 2015, from http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html

Kellmeyer, Phillipp, Ziegler, Wolfram, Peschke, Claudia, Juliane, Eisenberger, Schnell, Susanne, Baumgaertner, Annette, et al. (2013). Fronto-parietal dorsal and ventral pathways in the context of different linguistic manipulations. *Brain and Language*, *127*(2), 241–250.

Levy, Jonathan, Pernet, Cyril, Treserras, Sébastien, Boulanouar, Kader, Aubry, Florent, Démonet, Jean-Fronçois, & Celsis, Pierre. (2009). Testing for the dual-route cascade reading model in the brain: An fMRI Effective Connectivity Account of an Efficient Reading Style. *PLoS ONE*, *4*(8), e6675.

McKoon, Gail, & Ratcliff, Roger. (2007). Interactions of meaning and syntax: Implications for models of sentence comprehension. *Journal of Memory and Language*, *56*(2), 270–290.

Merolla, Paul A., Arthur, John V., Alvarez-Icaza, Rodrigo, Cassidy, Andrew S., Sawada, Jun, Akopyan, Filipp, et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, *345*(6197), 668–673.

Meystre, Stéphane M., Savova, Guergana K., Kipper-Schuler, Karin C., & Hurdle, John F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, *35*, 128–144.

Moon, Sungrim, Pahkhomov, Serguei, Liu, Nathan, Ryan, James O., & Melton, Genevieve B. (2014). A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, *21*(2), 299–307.

Price, Cathy J. (2013). Current themes in neuroimaging studies of reading. *Brain and Language*, *125*(2), 131–133.

Sowa, John F. (2010). Biological and psycholinguistic influences on architectures for natural language processing. Proceedings of the First Annual Meeting of the BICA Society, IOS Press, Incorporated, 221, 131.

Traxler, Matthew. (2012). Introduction to Psycholinguistics. Wiley-Blackwell.

Wu, Stephen T., Liu, Hongfang, Li, DDingcheng, Tao, Cui, Musen, Mark A., Chute, Christopher G., & Shah, Nigam H. (2012). Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *Journal of the American Medical Informatics Association*, *19*(e1), e149–56.

Yeatman, Jason D., Rauschecker, Andreas M., & Wandell, Brian A. (2013). Anatomy of the visual word form area: adjacent cortical circuits and long-range white matter connections. *Brain and Language*, *125*(2), 146–155.

Zhou, Wei, Torvik, Vetle I., & Smalheiser, Neil R. (2006). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, *22*(22), 2813–2818.