# Age of Exposure: A Model of Word Learning

**Mihai Dascalu[1], Danielle S. McNamara[2], Scott Crossley[3] and Stefan Trausan-Matu[1]**
[1]University Politehnica of Bucharest, 313 Splaiul Indepententei, Bucharest, Romania
[2]Arizona State University, PO Box 872111, Tempe, AZ 85287, USA
[3]Georgia State University, 25 Park Place, Ste 1500, Atlanta, GA 30303, USA
mihai.dascalu@cs.pub.ro, danielle.mcmamara@asu.edu, scrossley@gsu.edu, stefan.trausan@cs.pub.ro

## Abstract

Textual complexity is widely used to assess the difficulty of reading materials and writing quality in student essays. At a lexical level, word complexity can represent a building block for creating a comprehensive model of lexical networks that adequately estimates learners' understanding. In order to best capture how lexical associations are created between related concepts, we propose automated indices of word complexity based on Age of Exposure (AoE). AOE indices computationally model the lexical learning process as a function of a learner's experience with language. This study describes a proof of concept based on the on a large-scale learning corpus (i.e., TASA). The results indicate that AoE indices yield strong associations with human ratings of age of acquisition, word frequency, entropy, and human lexical response latencies providing evidence of convergent validity.

## Introduction

Measuring and quantifying the complexity of texts has been of particular interest in terms of aligning reading materials to a learner's level. However, determining a material's textual complexity is a difficult task as any potential measure is relative to the reader and individual differences that may arise due to prior knowledge, language familiarity or personal motivation and general interests. In addition to aligning reading material, as proposed in the Common Core State Standards Initiative (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), textual complexity has a strong role in the evaluations of students' readiness for college and their later careers (Dascalu, 2014).

The words that comprise a text are one textual constituent that contributes to the complexity of a text. Hence, measuring a text's difficulty relies on determining the underlying *complexity* of its *words*. One approach to estimating lexical complexity might depend on experts to annotate data and dictionaries and to create lists of words with their corresponding complexity. This qualitative method would provide important information, but it is not scalable and continuous updates may be required. In addition, the approach is not adaptable to different educational/domain contexts that can vary greatly. Therefore, automatic quantitative formulas of lexical complexity such as the Lexile reader measure (Stenner, 1996), the Flesch–Kincaid index (Kincaid, Fishburne, Rogers, & Chissom, 1975) or the Coleman-Liau index (Coleman & Liau, 1975) have become viable alternatives. These formulas provide automated estimates of text difficulty level based on the difficulty of the words and the sentences. Although these empirical and heuristic metrics do not provide perfect outputs, they work well in general and are widely adopted (about half of U.S. students from 3rd to 12th grade levels receive a Lexile measure each year; (Nelson, Perfetti, Liben, & Liben, 2012). However, a major problem with these metrics is that more in-depth discourse structures are not considered. For instance, Lexile measures are based on word frequencies and sentence length, whereas the Flesch-Kincaid index is built on average syllables per word. More modern research has indicated that lexical complexity is more complicated than these simple approaches and thus new approaches are necessary (Crossley, Greenfield, & McNamara, 2008). For instance, a word can have multiple meanings that are acquired gradually by students and allowing appropriate associations between concepts to be created over time. In order to model such a timeline, Landauer, Kireyev, and Panaccione (2011) adopted an approach for examining the evolution of a word's meaning through a large corpus analysis. They termed this approach "word maturity."

The aim of the research presented in this paper is to build on the word maturity metric proposed by Landauer et al. (2011) and create a model of lexical complexity that simulates the learning process as a function of experience with language. Therefore, in this study we present a model of lexical complexity that evaluates the learning curve of a word, the word's complexity, and the relations of its underlying meaning and associations to other concepts. Our new metric, named *Age of Exposure (AoE)*, describes a word's state of development in terms of inchoate, intermediate, and fully developed moments of acquisition which are modeled and generated from a learning corpus. Each intermediate model is built based on the learning materials available at a certain grade level (i.e., a *word's age of exposure*). The AoE metric thus reflects the age/grade level of a learner when s/he has been exposed to enough information to understand and create the appropriate associations for the given concept. Our approach also differs in that most computational research into lexical complexity has focused on using algebraic or probabilistic methods to represent words in semantic spaces starting from co-occurrence patterns of words in documents. Only a few approaches have considered the evolution of words' conceptualization and, in particular, the practical applications in Learning Analytics (LA). By training multiple semantic models on cumulative training data of increasing text difficulty, we are able to model the learning curve and potential difficulty of each word.

In this study, we present a proof of concept of our AoE model and corresponding complexity indices, as well as validations based on comparisons to word features such as human ratings of age of acquisition, word frequency, entropy, and human lexical response latencies. We find that the AoE indices strongly correlate with other indices of lexical complexity, thus providing evidence for the validity in measuring lexical complexity. Because the indices are automated, they can be used in a number of artificial intelligence systems to a) best align reading materials to learner's level of comprehension, b) improve the representation of concept maps of semantically related concepts filtered via complexity, or c) recommend students readings.

## State of the Art – Word Maturity

The *"word maturity"* metric estimates how a word's meanings are acquired gradually, depending on its complexity (Landauer et al., 2011). In extent, word maturity models the degree to which a word is known at different levels of language exposure. For example, children are generally exposed to a word such as *"dog"* early in life, leading to a greater likelihood of it being acquired. In contrast children are less likely to be exposed to a word such as *"focal"* at an early stage. Also, instead of considering words as independent concepts that can be either understood or not, a word maturity index sees them as bags of hidden associated concepts with each concept having a degree of contribution. One example of this is the word *"turkey"*. The word has a double meaning: the first and simplest meaning refers to the bird, but there is also a more complex meaning referring to "Turkey" – the country found in Europe. Each meaning has a different complexity level leading to the "country" concept most likely being acquired later in time.

Pearson Education currently uses word maturity to create assessment tools and personalized vocabulary instruction (http://www.readingmaturity.com/rmm-web/#/). Additional features such as word length, sentence length, within sentence punctuation, sentence and paragraph complexity, order of information and semantic coherence (within and between sentences; (Nelson et al., 2012) are also considered in the final assessment of a document's complexity. However, word maturity is proprietary and not all details for the computational analysis are available to the public, making comparisons difficult.

While not all implementation details of word maturity are available, the algorithm consists of the following principal stages (Kireyev & Landauer, 2011):

1. *Create an LSA (Landauer, Foltz, & Laham, 1998) space for each intermediate complexity model. Create an additional LSA space for the mature model*

For computing word maturity, multiple intermediate LSA spaces are built based on text corpora of increasing number of paragraphs. LSA is an unsupervised learning algorithm based on Singular Value Decomposition (SVD) often used in natural language processing to determine relations between documents and words while projecting them into a vector space. Given a document set, a term-document occurrence matrix $X$ is constructed using the word weights – usually Tf-Idf or log-entropy weights (Landauer, McNamara, Dennis, & Kintsch, 2007) – (rows) that appear in a document (the column). The obtained matrix is factorized into three matrices followed by a rank reduction. Given a fixed number $k$ ($k$<<rank of initial matrices - representing the most important latent dimensions), an approximation of $X$ is computed that has the property to have minimal errors in terms of the Frobenius norm. In the end, the semantic distance between words is computed as cosine similarity.

The collections of texts based on increasing paragraphs are used to represent cumulatively enlarged samples of documents generated by adding texts at successive Lexile levels. Therefore, more and more difficult texts are added to each intermediate complexity model, generating the mature corpus.

2. *Make the spaces compatible via Procustes Alignment (PA) in order to be able to compare concepts across different LSA spaces*

Comparing vectors obtained from different LSA spaces poses two problems: dimensionality and coordinate inconsistencies. Procustes Analysis (PA, (Krzanowski, 2000) can be used to align vectors in order to make them comparable. After applying PA, the comparison of word meanings in two different LSA spaces is reduced to analyzing the standard cosine metric for the two words in the joint space with compatible coordinates.

*3. Compute the word maturity as the similarity of a word in the target model with the same word in the mature model*

The previous alignment technique makes it possible to obtain a similarity metric between two words in separate corpora. At this point, each word's maturity level can be computed as the cosine similarity between the word in the intermediate model and the same word in the mature model. The visual representations of a word's evolution (Landauer et al., 2011) depends on the number of paragraphs within each training corpora, which can be also mapped on to the grade level of the underlying textual materials (Kireyev & Landauer, 2011). To validate the maturity function, a time-to-maturity index is defined as the minimum at which the word maturity (or semantic similarity to the mature space) reaches a particular threshold α. When compared to human vocabulary development word lists (e.g., Age Of Acquisition Norms), the Time-to-Maturity (α = 0.45) index had a Spearman correlation of *.72* to the (Gilhooly & Logie, 1980) AoA norms (n = 1643) and a *.64* correlation to Bristol (Stadthagen-Gonzalez & Davis, 2006) AoA norms (n = 1402).

# Method

Similar to Landauer et al. (2011), our aim is to create a generalized model of word complexity - Age of Exposure - that simulates the potential learning curve of a concept based on its associations to other words. One main difference is that we use Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to accomplish this rather than LSA. Our method consists of the following steps:

- Create incremental corpora in terms of dimensions and of complexity in order to model the word learning process. The largest/highest ranking space is considered the most developed or mature space.
- Train dedicated LDA models for each intermediate corpus and on the most developed corpus.
- Align and match intermediate model topics to the most developed model by creating a bi-partite graph and by applying a flow algorithm.
- Based on the matching, compare the representation of a concept in an intermediate model with the matched topic distribution in the aligned most developed model via cosine similarity.

This approach provides a series of matchings between each incremental corpus and the most developed/mature corpus, denoting the representation of a concept in terms of topic distributions in incremental semantic models. Based on these series of [0 - 1] similarity values, we develop different Age of Exposure (AoE) indices, presented later in detail, in order to obtain an estimation of each word's complexity.

## Building LDA intermediate and mature models

AoE relies on Latent Dirichlet Allocation (LDA), a reliable topic modeling techniques that uses a generative probabilistic process to infer underlying topics (Blei et al., 2003). Starting from the presumption that documents integrate multiple topics, each document can now be considered a random mixture of corpus-wide topics. A topic is a Dirichlet distribution (Kotz, Balakrishnan, & Johnson, 2000) over the vocabulary simplex (the space of all possible distributions of words from the training corpora) in which semantically related words have similar probabilities of occurrence.

LDA captures and creates word associations based on co-occurrence data, which means that an in-depth view of word understanding is provided, based on its links to other semantically related concepts. Using LDA provides extendibility and a wider applicability because a major constraint of LSA is that the SVD decomposition is a highly computational and resource demanding process, while LDA can be easily applied on larger corpora. Moreover, LDA partially addresses polysemy (i.e., ambiguity in the senses attributed to a word), which is disregarded in LSA. Polysemy is reflected in the number of different topics/contexts containing the given word. In addition, the usage of LDA is more straightforward and better highlights the associations between concepts through the underlying latent topics.

Although LDA has proven to be reliable in extracting topics and has the lowest perplexity when compared to other probabilistic semantic models (Blei et al., 2003), we must also consider its drawbacks. First, there are no actual significances assigned to topics as words have corresponding probabilities, but there is no overarching domain classification marking a certain topic (e.g., the is no explicit marking for frequently encountered topics on “religion”, “war”, “economy”, etc.). Moreover, topics are not equiprobable (Arora & Ravindran, 2008) and there is no imposed ordering (e.g., topic *i* in one training is centered on concepts related to “politics”, but in another it can consist of words from any other semantically related subset of concepts). Second, there are inevitable estimation errors by using an approximate inference model, which are more notable when addressing smaller texts with a more uncertain mixture of topics. Third, like LSA, LDA is blind to word order, but polysemy is now reflected in the membership of the same word, with high probabilities, in multiple topics. Lastly, LDA loses LSA's cognitive significance from a psychological point of view (Landauer et al., 2007).

## Topic matching

Applying LDA to the intermediate and mature corpora creates increasingly complex LDA models, but comparing topics across LDA spaces is not directly possible as the latent variable between LDA spaces are independent and have no direct linkage between models. Therefore, our aim is to track, match, and align the topics behind LDA spaces. However, this task of topic alignment can be reformulated to a more popular problem: finding the best node matching in a bipartite graph while minimizing the total cost (see Figure 1). Therefore, every topic from an intermediate corpus can be matched to a topic in the most developed model with a cost, or more specifically, the Jensen-Shannon divergence (JSD) between all word probabilities from the two topics (Dascalu, 2014). Computing the minimum cut with the minimum total cost can be performed by using the maximum flow (Ford–Fulkerson) algorithm with a few changes. First, after adding a virtual source and a virtual sink with infinite capacities (see Figure 1), the Bellman-Ford algorithm was used (Cormen, Leiserson, Rivest, & Stein, 2009) to determine the optimal path via intermediate and mature topics because the semantic similarity function (JSD) can have negative costs. Second, only pairs of topics that correspond to the same or highly similar meanings were selected, ensuring that each topic from an intermediate model had a corresponding topic in the mature one.

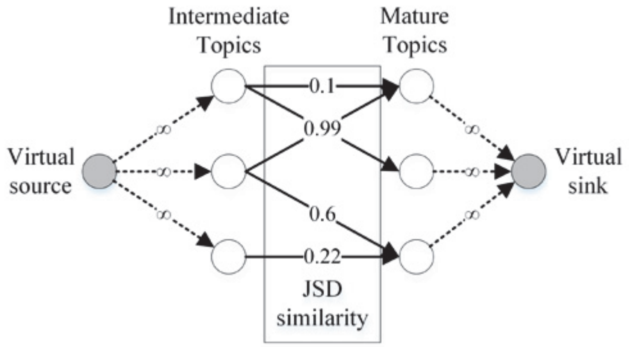

*Figure 1. Bipartite graph depicting the potential matching based on JSD between topics, as well as the virtual source and sink used for the maximum flow algorithm.*

Overall, the topic matching is a completely automated process that builds the bipartite graph on the JSD between all word probabilities from two topics (one from the intermediate model and the other from the mature model). Similar to Landauer's word maturity model, there is no automatic alignment between two different LSA spaces (coordinate inconsistency of each $i^{th}$ dimension of the $k$ existing ones between the two models), which was addressed in their case via Procustes Alignment.

## Computing AoE

At this point in the computation of AoE, multiple LDA intermediate and mature spaces are trained allowing the topics of any intermediate models to be matched with mature models. Moreover, each word from any specific corpus has a discrete probability distribution over the set of topics corresponding to that LDA model. Based on the previous matching, a permutation of the topics is performed enabling the comparison of a word's representation in an intermediate LDA model to its aligned topic distribution in the mature space. Therefore, our AoE function per intermediate model is captured as the cosine similarity between the word in the intermediate space versus the word's topic distribution in the mature model. In other words, AoE gradually captures the degree to which a word is correctly represented with regards to the emergent latent topics or the level of its potential understanding in any intermediate model.

By considering multiple snapshots derived from intermediate LDA models, we obtain the potential learning curve of a concept by simulating the creation of a learner's word associations based on the provided corpora. Afterwards, multiple AoE indices can be developed that captured a word's complexity level based on the adequacy of its associations. Of these potential indices, we focus on the following:

- *Inverse Average Similarity* = 1 – Avg (similarity values per intermediate model): The easier a concept is, the faster it is will be represented correctly in an intermediate model versus the most developed space.
- *Inverse Linear Regression Slope*: A linear regression from (0, 0) up to (1, 1) was generated as all intermediate cosine values are equally distributed on the Ox axis. AoE is estimated as the inverse of the slope (AoE = 1/slope).
- *Index above Threshold*: The index of the intermediate model for which the cosine value exceeds an imposed threshold (experimentally, a threshold of .4 provided the best results when considering all possible thresholds from 0.4 to 0.7 with a 0.1 increment).
- *Index Polynomial Fit above Threshold*: Because individual word development is reflected by the amount of simulated reading (the volume of text that a learner is potentially exposed to, up until a certain grade level), the polynomial fit of degree 3 provides a continuous measure and follows more smoothly the similarities than a linear interpolation or a 2nd degree fit. The index represents the first intermediate model/grade level that exceeds an imposed threshold (a 0.4 threshold provided the best results).
- *Inflection Point of the Polynomial Fit*: After creating an extended data series that ensures the *S* shape of the interpolation for reducing noisy input, the inflection point of the generated polynomial fit best marks the beginning of correctly representing a given concept based on its associations.

Although tightly connected, the provided indices for AoE capture different specificities and were specifically designed to capture different traits of the modeled learning curve (slope, inflection point, and increase over an imposed threshold).

## Training Corpora

Our AoE indices were implemented using the TASA corpus (http://lsa.colorado.edu/spaces.html) which was segmented based on Degrees of Reading Power (DRP; (Koslin et al., 1987) into 13 grade levels (McNamara, Graesser, & Louwerse, 2012). The *n*-th AoE intermediate model contained all the documents of complexity *1* up to *n* (with a corresponding notation of [1 – n]). Lemmatization and stop words elimination were used when preparing the corpus. The TASA corpus is used to create a proof of concept. However, our model can be easily applied on any other textual databases in order to compute both domain specific and domain general AoEs.

Text preprocessing was performed, but the preprocessing phases described later on represent only an optional refinement of our model. Lemmatization was performed by applying the Stanford Core NLP MorphAnnotator (http://nlp.stanford.edu/software/) and our stop word list was a slightly modified version of the Snowball list (http://snowball.tartarus.org). We hypothesized that part of speech (POS) tags would be detrimental to the overall performance of the AoE indices because they might generate conflicting AoE scores for each POS tag. In other words, a concept or lemma would develop an AoE score for each POS tag. We thus did not use POS tagging on the corpus.

Because we were dealing with incremental corpora and the sizes followed an almost arithmetic progression (see Table 1), the trained LDA models used for this proof of concept had a similar progression in terms of number of topics. The arithmetic progression from Table 1 yield better results in terms of correlations to other word features than a proportional growth with the number of types. For the most developed model, we opted to use 100 topics as indicated by Blei (2012). Further refinements of AoE will consider optimizing the identification of topics via Hierarchical Dirichlet Processes (HDP; Teh, Jordan, Beal, & Blei, 2006) used to infer the number of topics from each intermediate or mature model.

Because low ranking intermediate models consisted only of general words contained within a number of mature topics, the actual perception between adjacent intermediate models was disrupted. Moreover, as there were no actual best matches between low-ranking intermediate topics and the mature ones due to the limited vocabulary and corpora, the partial associations that could have been generated by other concepts degenerated the AoE results. In addition, as a particular implementation tweak, we had a cosine measure between two vectors containing just one value in some particular cases, which is quite normal because a word in an intermediate space could be present in only one topic. Due to the fact that cosine similarity would automatically create an optimal matching of 1, the individual similarity value was replaced with the ratio of word weights between the models.

*Table 1. Statistics of intermediate and mature models after lemmatization and stop-words elimination.*

| Grade level | Types | Tokens | Paragraphs | Topics |
|---|---|---|---|---|
| [1 – 1] | 8,377 | 367,277 | 3,612 | 5 |
| [1 – 2] | 12,601 | 681,087 | 6,530 | 10 |
| [1 – 3] | 15,652 | 962,751 | 9,078 | 15 |
| [1 – 4] | 18,492 | 1,292,570 | 12,022 | 20 |
| [1 – 5] | 22,457 | 1,841,657 | 16,810 | 25 |
| [1 – 6] | 25,930 | 2,432,460 | 21,824 | 30 |
| [1 – 7] | 26,976 | 2,620,402 | 23,378 | 35 |
| [1 – 8] | 27,967 | 2,807,591 | 24,912 | 40 |
| [1 – 9] | 29,057 | 3,004,454 | 26,499 | 45 |
| [1 – 10] | 30,909 | 3,378,804 | 29,465 | 50 |
| [1 – 11] | 32,553 | 3,728,315 | 32,160 | 55 |
| [1 – 12] | 33,268 | 3,892,696 | 33,409 | 60 |
| Mature model | 37,633 | 5,084,243 | 41,866 | 100 |

In total, 26,470 words were represented in our AoE model. Around 7,000 words from the [1-12] intermediate model had singular occurrences and were not representative within the LDA spaces. Out of the 26,470 words included in the AoE model, 6,258 words (23.64%) had an approximate monotonic growth. That is to say, for each intermediate model *i*, the current cosine similarity to the mature model was greater than 75% of the previous (*i-1*) similarity. This approximation of a trending growth for word assimilation in all subsequent models was frequently encountered for concepts that had only one sense. In contrast, polysemy is reflected in decreases and spikes in the learning curve of a concept as new senses are introduced into more complex models (e.g., see pattern for “class” in Figure 2). Simple concepts were well associated from the beginning (e.g., “chocolate”, “happy”), whereas more complex words become well represented later on (e.g., “tech” or “clustering”). Specific scientific concepts were only introduced in higher grade levels (e.g. “virus”) and their conceptualization may never become fully adequate without the entire mature corpora (e.g., “singularity”).

# Results

To validate our AoE indices, we took a convergent validity approach in which we assessed the degree to which our AoE indices converged (i.e., correlated) with other word features to which they should be theoretically similar. Specifically, we selected word features related to age of acquisition, word frequency, word entropy, response latencies, and word familiarity. All of these variables are strongly related to lexical sophistication and/or lexical acquisition. The selected indices are reported by the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) (Kyle & Crossley, 2015) and briefly discussed below.

*Table 2. Correlations between AoE indices and convergent variables.*

| Selected index | Inverse Average Similarity | Inverse Linear Regression Slope | Index Above Threshold | *Index Polynomial Fit Above Threshold* | *Inflection Point of the Polynomial Fit* |
|---|---|---|---|---|---|
| Kuperman AoA | .884 | .716 | .912 | .891 | .893 |
| SUBTLEXus word frequency | -.742 | -.599 | -.765 | -.749 | -.774 |
| SUBTLEXus entropy | -.752 | -.615 | -.776 | -.761 | -.780 |
| Word naming latency | .761 | .611 | .779 | .761 | .774 |
| Lexical decision latency | .754 | .616 | .766 | .756 | .753 |

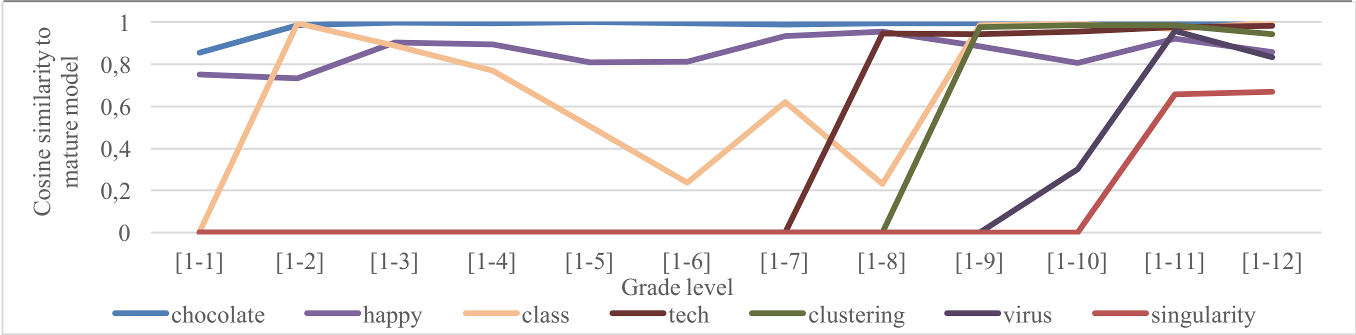

*Figure 2. Comparative view of AoE for selected words.*

- *Age of Acquisition (AoA):* (AoA) indices are based on human judgments of the age that a particular word is learned (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). We selected AoA scores for 30,121 lemmas based on Kuperman et al. (2012).
- *Word Frequency:* Frequency effects are based on the notion that words that are more frequent in natural language data are learned earlier and used more often than words that are less frequent in natural language data. We selected frequency indices from the SUBTLEXus corpus (Brysbaert & New, 2009), which includes 51 million words.
- *Word Entropy:* Entropy measures account for how widely a word or word family is used, usually by providing a count of the number of documents in which that word occurs. We selected entropy indices based on SUBTLEXus, which reports entropy based on 8,388 texts.
- *Response Latencies:* Lexical decision and naming response times were obtained from The English Lexicon Project (ELP) (Balota et al., 2007). This dataset includes response latencies for 80,962 real word and nonword stimuli (40,481 each), including both mono- and multisyllabic words.

Pearson product moment correlations (see Table 2) demonstrated that all AoE variables had strong (i.e., strong effects, $r > .500$) and significant relations (i.e., $p < .001$) with the selected convergent validity indices related to lexical sophistication and knowledge. Moreover, the correlations surpass the publicly available results for the word maturity measure.

## Conclusions

Our AoE metric is a reproducible and scalable model that can be easily applied on different textual corpora and databases. In contrast to word maturity, AoE has introduced more indices for estimating word complexity that better correlate to human ratings. Moreover, word maturity introduces an additional approximation for producing a calibrated scale. With regards to the latter step, AoE is more straightforward and more accurate, providing an in-depth perspective of complexity by simulating word learning based on potential associations created across time.

A downside of our proof-of concept AoE model is our use of the TASA corpus as a proxy for world experience, based on DRP scores (Carver, 1985). As such, our intermediate models are relatively artificial. We considered merging adjacent levels and creating a stronger baseline (e.g., a larger 1st complexity model using 1-2 or 1-3 levels) to generate smother learning curves, but we opted to present the most granular results possible in order to better emphasize the benefits of our AoE model versus word maturity. Moreover, we are fully aware of a potential circular argument; a more adequate corpus would contain actual texts extracted from schoolbooks corresponding to each grade level. Nevertheless, this study provides a successful proof of concept that automatic textual complexity assessment based on simple surface measures can be used to create an initial segmentation of the training corpora.

However, by using better corpora and by refining the number of imposed topics per model, we should be able to train specialized AoE models in future iterations of this work. We envision that AoE indices can be used to better match texts to readers, to better analyze complexity in text and speech, and to provide better feedback to users in intelligent systems.

## Acknowledgments

This research was partially supported by the 2008-212578 LTfLL FP7 and POSDRU/159/1.5/S/134398 projects, as well as by the NSF grants 1417997 and 1418378 to Arizona State University. We would also like to thank Lucia Larise Stavarache, Philippe Dessus, Teodor Rosu and Laura Allen for their involvement in the design of AoE.

## References

Arora, R., & Ravindran, B. 2008. Latent dirichlet allocation based multi-document summarization. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*, 91–97. Singapore: ACM.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., . . . Treiman, R. 2007. The English Lexicon Project. *Behavior Research Methods, 39*(3): 445–459.

Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM, 55*(4): 77–84.

Blei, D.M., Ng, A.Y., & Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5): 993–1022.

Brysbaert, M., & New, B. 2009. Moving beyond Kucera 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.

Carver, R.P. 1985. Measuring Readability using DRP Units. *Journal of Literacy Research, 17*(4): 303–316.

Coleman, Meri, & Liau, T.L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*: 283–284.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C. (Eds.) 2009. *Introduction to Algorithms* (3rd ed.). Cambridge, MA: MIT Press.

Crossley, S.A., Greenfield, J., & McNamara, D. S. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*: 475–493.

Dascalu, M. 2014. *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* Vol. 534. Switzerland: Springer.

Gilhooly, Ken J., & Logie, R. H. 1980. Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 word. *Behaviour Research Methods & Instrumentation, 12*: 395–427.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. 1975. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Air Station Memphis: Chief of Naval Technical Training,.

Kireyev, K., & Landauer, T.K. 2011. Word Maturity: Computational Modeling of Word Knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 299–308. Portland, Oregon: ACL.

Koslin, B.L., Zeno, S.M., Koslin, S., Wainer, H., & Ivens, S.H. 1987. *The DRP: An effectiveness measure in reading*. New York, NY: College Entrance Examination Board.

Kotz, S., Balakrishnan, N., & Johnson, N.L. (2000). Dirichlet and Inverted Dirichlet Distributions *Continuous Multivariate Distributions* (Vol. 1, pp. 485–527). New York, NY: Wiley.

Krzanowski, W.J. 2000. *Principles of Multivariate Analysis: A User's Perspective*. Oxford: Oxford University Press.

Kuperman, Victor, Stadthagen-Gonzalez, Hans, & Brysbaert, Marc 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4): 978–990.

Kyle, K., & Crossley, S.A. in press. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*.

Landauer, T.K., Foltz, P.W., & Laham, D. 1998. An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2/3): 259–284.

Landauer, T.K., Kireyev, K., & Panaccione, C. 2011. Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1): 92–108.

Landauer, T.K., McNamara, D.S., Dennis, S., & Kintsch, W. (Eds.) 2007. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

McNamara, D.S., Graesser, A.C., & Louwerse, M.M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini, E. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham, MD: R&L Education.

National Governors Association Center for Best Practices & Council of Chief State School Officers 2010. Common Core State Standards, Washington D.C.: Authors.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance, Council of Chief State School Officers, Washington, DC.

Stadthagen-Gonzalez, Hans, & Davis, C. J. 2006. The Bristol Norms for Age of Acquisition, Imageability and Familiarity. *Behavior Research Methods, 38*: 598–605.

Stenner, A.J. 1996. Measuring reading comprehension with the Lexile Framework, MetaMetrics, Inc., Durham, NC.

Teh, Y.W., Jordan, M. I., Beal, M.J., & Blei, D.M. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association, 101*: 1566–1581.