Psycholinguistic Models of Sentence Processing Improve Sentence Readability Ranking

David M. Howcroft[1] and **Vera Demberg**[1,2]
[1] Department of Language Science and Technology
[2] Department of Computer Science
Saarland Informatics Campus, Saarland University
Saarbrücken, Germany
{howcroft,vera}@coli.uni-saarland.de

Abstract

While previous research on readability has typically focused on document-level measures, recent work in areas such as natural language generation has pointed out the need of sentence-level readability measures. Much of psycholinguistics has focused for many years on processing measures that provide difficulty estimates on a word-by-word basis. However, these psycholinguistic measures have not yet been tested on sentence readability ranking tasks. In this paper, we use four psycholinguistic measures: idea density, surprisal, integration cost, and embedding depth to test whether these features are predictive of readability levels. We find that psycholinguistic features significantly improve performance by up to 3 percentage points over a standard document-level readability metric baseline.

1 Introduction

Previous work on readability has classified or ranked texts based on document-level measures such as word length, sentence length, number of different phrasal categories & parse tree depth (Petersen, 2007), and discourse coherence (Graesser et al., 2004), inter alia. However, not all applications that need readability ratings deal with long documents. For many applications in text simplification, computer-aided language learning (CALL) systems, authorship tools, translation, and information retrieval, sentence-level readability metrics are direly needed.

For instance, an automatic text simplification system must begin by asking which portions of a text need to be simplified. To this end, a measure that can assign ratings on a sentence-by-sentence level can help target simplification only to those sentences which need it most, and such measures also serve to confirm that the resulting 'simplified' sentence is in fact simpler than the original sentence.

Similarly, CALL and other pedagogical systems will benefit if it is possible to predict which portions of a text will be harder for students. Authorship tools can offer more specific editorial advice when they know why individual sentences can cause difficulties for readers. Translation tools can aim to preserve not just meaning but also the approximate difficulty of the sentences they are translating or use a sentence-level difficulty metric to target output that is easier to understand. Furthermore, information retrieval systems also benefit when they can return not merely relevant texts, but also texts appropriate to the reading level of the user. Recently there has been an increased interest in sentential models of text difficulty in the automatic text simplification and summarization communities in particular (Vajjala and Meurers, 2014; Macdonald and Siddharthan, 2016).

One area that has produced a lot of research on sentence level processing difficulty is psycholinguistics. Over the past three decades, a number of theories of *human* sentence processing (i.e. reading) have been proposed and validated in a large variety of experimental studies. The most important sentence processing theories have furthermore been implemented based on broad-coverage tools, so that estimates for arbitrary sentences can be generated automatically. For example, eye-tracking studies of reading times on a large corpus of newspaper text have found that measures such as *integration cost* and *surprisal* provide partial explanations for subjects' reading behavior (Demberg and Keller, 2008).

This paper leverages these implemented measures based on psycholinguistic theories of sen-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 958–968,
Valencia, Spain, April 3-7, 2017.

tence processing in order to test whether they can help to more accurately score individual sentences with respect to their difficulty. In the process, we evaluate the contributions of the individual features to our models, testing their utility in examining fine-grained distinctions in sentence difficulty. Section 2 reviews the literature on readability in general before we shift to psycholinguistic theories of sentence processing in Section 4. In Section 5 we discuss our methods, including the corpora used, how features were extracted, and the set up for our averaged perceptron models. Section 6 presents our findings which we connect to related work on sentence-level readability models in 3. Finally we offer our conclusions and suggestions for future work in Section 7.

2 Readability

Chall's (1958) comprehensive review of readability research in the first half of the 20th century divides the early work in readability into "survey and experimental studies" and "quantitative associational studies". Studies of the former category took place during the 1930s and 1940s and included surveys of expert and reader opinion as well as experimental studies which manipulated texts according to one variable at a time in order to determine the effects of those variables on readers. The results of these studies suggest that, once you have managed to control for reader interest in the content of a text, the most important factor with respect to its readability is its 'style', e.g. its "scope of vocabulary and...kinds of sentences" (Gray and Leary, 1935, as quoted in (Chall, 1958)).

Our study belongs to the second class, relating the features of a text to its ordering relative to some other texts. The earliest work in this direction was by L. A. Sherman, who proposed a quantitative analysis of text difficulty based on the number of clauses per sentence, among other features (Sherman, 1893). Where Sherman's pedagogical focus was on literature, Lively & Pressey (1923) focused on vocabulary as a bottleneck in science education. Work in this vein led to the development of a number of readability formulae in the mid-20th century[1], including the familiar Flesch-Kincaid Grade-Level score (Kincaid et al., 1975).

These formulae typically use a linear combination of average word length and average sentence length, though some also incorporate a vocabulary-diversity term. The simple, two-feature versions of these models are still widely used, and inspired our BASELINE model.

More recently, Petersen (2007) sought to apply familiar natural language processing techniques to the problem of identifying text difficulty for non-native readers. In particular, she used a number of parse-based features which captured, for example, the average number of noun and verb phrases per sentence and the height of the parse tree. Petersen trained SVM classifiers to classify texts as belonging to one of four primary school grade levels based on the Weekly Reader educational newspaper[2]. These document-level models achieved F-scores in the range of 0.5 to 0.7, compared to the F-scores between 0.25 and 0.45 achieved by the Flesch-Kincaid Reading Ease score for the same texts.

Recent work has also looked at features related to discourse and working memory constraints. Feng et al. (2009) worked on a model of readability for adults with intellectual disabilities. Considering working memory constraints, they extracted features related to the number of entities mentioned in a document and the 'lexical chains' (Galley and McKeown, 2003) that connected them. They found that their features resulted in a better correlation (Pearson's $r = -0.352$) compared to both Flesch-Kincaid score ($r = -0.270$) and a number of 'basic' linguistic features based on those used by Petersen & Ostendorf (2009) ($r = -0.283$).[3]

Coh-Metrix (Graesser et al., 2004) also includes a number of measures related to discourse coherence, for example. Such features are not suited to the problem of determining the difficulty of sentences in isolation, but they have also been shown to better predict readability for second-language learners compared to 'traditional' readability measures like those described above (Crossley et al., 2011).

[1] For a comprehensive review of the literature up to 1958, we recommend (Chall, 1958). For a more recent review of the literature, we recommend Chapter 2 of (Vajjala, 2015). For an introduction to some of the major studies from the 20th century, we recommend the self-published (Dubay, 2007).

[2] http://www.weeklyreader.com

[3] Correlations here are negative because Feng et al. correlated predicted reading levels with the performance of adults with intellectual disabilities on comprehension tests. The adults with disabilities are expected to perform worse on the comprehension test as the grade level of the text increases.

3 Measuring Sentence Complexity

Classification Few studies to date have addressed sentence-level readability for English. Napoles & Dredze (2010) built their own corpus with documents from English and Simple English Wikipedia to train both document- and sentence-level classifiers. Using bag-of-words features, unigram and bigram part-of-speech features, type-token ratio, the proportion of words appearing on a list of easier words, and parse features similar to Petersen's, their binary classifier achieved an accuracy of 80.8% on this task. The structure of this task, however, is not suited to text simplification applications, because the sentences are not controlled for meaning. Classifying a sentence in isolation as more likely to be from Simple Wikipedia or English Wikipedia is not as useful as a model trained to differentiate sentences carrying the same meaning. This work is not directly comparable to that of Vajjala & Meurers (2012; 2014) or subsequent work on ranking sentences by their complexity due to the differences in choice of corpora and task structure.

In the medical domain, Kauchak et al. (2014) also looked at sentence-level classification, identifying sentences as being either simple or difficult. Their features included word length, sentences length, part-of-speech counts, average unigram frequencies and standard deviation, and the proportion of words not on a list of the five thousand most frequent words as well as three domain-specific features based on an ontology of medical terminology.

Ranking Vajjala & Meurers (Vajjala and Meurers, 2014; Vajjala, 2015) were the first to look at *ranking* sentences rather than classifying them, having observed that the distributions of predicted reading levels across the two subcorpora of the Parallel Wikipedia corpus (Zhu et al., 2010, PWKP) were different. While the Simple English portion of the corpus was clearly skewed toward the lower grade levels, it appears that the English portion of the corpus was evenly distributed across all grade levels, making binary-classification difficult.

This led Vajjala & Meurers to develop a ranking model using the predicted reading levels from a multiclass classifier trained on whole documents. For each sentence pair, they assumed that the English Wikipedia sentence should be classified at a higher level than the Simple English Wikipedia sentence. Using a hard cut-off (i.e. $rank(sent_{english}) > rank(sent_{simple})$), their model achieved about 59% accuracy, although this improved to 70% by relaxing the inequality constraint to include equality. Based on the finding that 30% of sentence pairs from the PWKP corpus are incorrectly ranked despite lying within one reading level of each other, we hypothesize that finer-grained distinctions may be necessary to tease apart the differences in related pairs of sentences.

Offline Psycholinguistic Features While Vajjala & Meurers (2012; 2014) do use some psycholinguistically-motivated features, their features are primarily lexical in nature and therefore complementary to ours, which depend on the sentence processing context. They drew psycholinguistic features from the MRC psycholinguistic database (Wilson, 1988), including word familiarity, concreteness, imageability, meaningfulness, and age of acquisition. These features were coupled with a second age of acquisition database and values related to the average number of senses per word.

Towards online considerations More recently, Ambati et al. (2016) used an incremental parser to extend Vajjala & Meurers work. Since human processing is incremental, they reasoned, features from an incremental parser might be more informative than features extracted from a non-incremental parser. To this end, they used the incremental derivations from a combinatory categorial grammar (CCG) parser. Ambati et al. ran several models on the English and Simple English Wikipedia data set (Hwang et al., 2015, ESEW):one using only the syntactic features from (Vajjala and Meurers, 2014); another (INCCCG) using only features from the incremental parser; and INCCCG+, incorporating morpho-syntactic and psycholinguistic features from (Vajjala and Meurers, 2014). At the sentence level, they include sentence length, number of CCG constituents in the final parse, and the depth of the CCG derivation. They also use count features for the number of times each CCG derivation rule is applied (e.g. forward application, type-raising). Finally, they include counts of different CCG syntactic categories as well as the average 'complexity' of the syntactic categories. While the parser they use is inspired by human behavior, in that it is an incremental parser, these features do not re-

late to any specific linguistic theory of sentence processing.

The work presented here is most comparable to that of Vajjala & Meurers and Ambati et al., as we all address the problem of ranking sentences according to their linguistic complexity. Our study is the only one of the three to examine features based on theories of online sentence processing. Ambati et al. (2016) provide accuracy information for their own features as well as Vajjala & Meurers' (2014) features on the English and Simple English Wikipedia corpus (ESEW) which we use, but used a 60-20-20 training-dev-test split where we used 10-fold cross-validation, making the results not directly comparable.

4 Theories of Online Sentence Processing

For our purposes, we focus on readability as reading ease and on linguistic constraints in particular, rather than constraints of medium (relating to e.g. legibility), reader interest, or comprehensibility. Without directly modeling comprehensibility, we assume that making material easier to read will also make it easier to comprehend. Here we focus on four psycholinguistic theories of human sentence processing: idea density, surprisal, integration cost, and embedding depth.

Kintsch (1972) defined **propositional idea density** as the ratio of propositions or ideas to words in the sentences.[4] Keenan & Kintsch conducted two different experiments in order to examine free reading behavior as well as subjects' performance in speeded reading conditions. They found that "the number of propositions [in a text] had a large effect upon reading times, [but] it could only account for 21% of their variance" when subjects were allowed to read freely. Subjects' overall recall was worse for more dense texts in the speeded reading condition. In addition to effects of idea density, they found that propositions which were presented as surface-form modifiers (as opposed to, e.g., main verbs) were "very poorly recalled" and that propositions playing a subordinate role relative to another proposition were also less-well recalled. Finally, propositions involving a proper name were generally recalled better than similar propositions involving, e.g., a common noun.

While Kintsch & Keenan (1973) looked at the influence of propositional idea density on reading times and recall for both individual sentences as well as short paragraphs, work since the 1970s has been limited to the level of multiple sentences and used primarily as an indicator of cognitive deficits (Ferguson et al., 2014; Bryant et al., 2013; Farias et al., 2012; Riley et al., 2005). This paper returns to the examination of idea density's applicability for individual sentences.

Surprisal, on the other hand, has been widely examined in theories of language comprehension at a variety of levels, including the word- and sentence-levels. **Surprisal** is another word for Shannon (1948) information, operationalized in linguistics as the probability of the current word conditioned on the preceding sequence of words:

$$\textit{surprisal}(w_n) = -log(P(w_n|w_1 \dots w_{n-1})) \quad (1)$$

where w_i is the i^{th} word in the sentence and $P(w_1 \dots w_i)$ denotes the probability of the sequence of i words $w_1 \dots w_i$.

One reason psycholinguists consider surprisal as a factor in sentence processing difficulty is that it makes sense in a model of language users as rational learners. Levy (2008) argues the rational reader's attention must be spread across all possible analyses for the sentence being observed. Based on prior experience, the reader expects some analyses to be more probable than others and therefore allocates more resources to those analyses. In this analysis, surprisal is derived as a measure of the cost paid when the reader misallocates resources: when a new word invalidates a highly probable analysis, the reader has effectively 'wasted' whatever resources were allocated to that analysis. The notion of surprisal is also used in theories of language production, see the Uniform Information Density hypothesis (Jaeger, 2006; Levy and Jaeger, 2007; Jaeger, 2010, UID).

While surprisal focuses on predictability effects in sentence processing, Gibson's (1998; 2000) **Dependency Locality Theory** (DLT) focuses on the memory cost of recalling referents and integrating new ones into a mental representation. DLT proposes that the the distance between syntactic heads and dependents, measured by the number of intervening discourse referents, approximates the difficulty that the listener or reader will have integrating the two units. This model maintains that the act of creating a new discourse referent and holding it in memory makes it more difficult to recall

[4] This notion of idea density is closely related to Perfetti's (1969) notion of *lexical density* insofar as both are related to the number of so-called *content words* in the text.

a previous discourse referent and connect that discourse referent to the current one.[5]

In addition to *integration cost*, DLT proposes a *storage cost* associated with the number of open dependencies that must be maintained in memory. The notion of connected components in van Schijndel et al.'s (2012; 2013) incremental parsing model picks up this idea. Related models were also suggested earlier by Yngve (1960) and Miller's (1956a; 1956b) whose work was based on results showing that human working memory is limited to 7 ± 2 items. Yngve's mechanistic, incremental model of language production considered the evaluation of phrase structure grammars (PSGs) in a system with finite memory, exploring the structure speakers must keep track of during production and how grammars might be structured to avoid overtaxing working memory.

Van Schijndel et al. develop this idea further in the context of a hierarchical sequence model of parsing. In this incremental model of parsing, at each stage the reader has an *active* state (e.g. S for sentence) and an *awaited* state (e.g. VP for verb phrase).[6] At each new word, the parser must decide between continuing to analyze the current connected component or hypothesizing the start of a new one.[7]

These measures provide an idealized representation of the number of different states a human parser must keep track of at any point in time. We refer to this number of states as the **embedding depth** of a sentence at a particular word, and the `ModelBlocks` parser of van Schijndel et al. (2012) calculates this number of states averaged over the beam of currently plausible parses. Also of interest is the *embedding difference*, which is the embedding depth at the present word relative to the previous word, elaborated upon in the following example.

Consider the state described above (i.e. that of being in the active state S and awaiting state VP) might be reached after a reader has observed a noun phrase, resulting in the state S/VP. This means that the word sequence observed so far will be consistent with a sentence if the reader now observes a verb phrase. If, however, the next word in the input is inconsistent with the start of a verb phrase (e.g. the relative clause marker *that*), then this parse will be ruled out and another must be considered. At this point the parser must hypothesize the beginning of a new connected component, i.e. a new syntactic substructure that must be completed before continuing to parse the top-level of the sentence. Therefore, the parser must now keep track of two states: (1) the fact that we are still looking for a VP to complete the overall sentence; and (2) the fact that we now have a relative clause to parse before we can complete the current NP. In this example, we are at embedding depth 1 or 0 up until we encounter the word *that*, which increases the embedding depth by 1, resulting in a nonzero embedding difference score.

4.1 Experimental Evidence

We have already explained the experimental findings of Kintsch & Keenan (1973) with respect to idea density, but what behavioral evidence is there to suggest that the remaining theories are valid?

Demberg & Keller (2008) examined the relationship between both surprisal and integration cost and eye-tracking times in the Dundee corpus (Kennedy and Pynte, 2005) Demberg & Keller found that increased surprisal significantly correlated with reading times. Although they found that integration cost did not significantly contribute to predicting eye-tracking reading times in general, its contribution was significant when restricted to nouns and verbs. They also found that surprisal and integration cost were uncorrelated, suggesting that they should be considered complementary factors in a model of reading times. Another eye-tracking study divided surprisal into lexical and synactic components, finding that lexical surprisal was a significant factor but not syntactic surprisal (Roark et al., 2009).

Wu et al. (2010) examined surprisal, entropy reduction, and embedding depth in a study of psycholinguistic complexity metrics. Their study of the reading times of 23 native English speakers reading four narratives indicated that embedding difference was a significant predictor of reading times for closed class words. Moreover, this contribution was independent of the contribution of surprisal, indicating that the two measures are

[5] Gildea & Temperley (2010) measure dependencies in terms of word span, such that adjacent words have a dependency length of one. This approach produces similar difficulty estimates nouns and verbs, with the caveat that distances are systematically increased, and is defined for all words in a sentence.

[6] In Combinatory Categorial Grammar notation, this state is denoted S/VP.

[7] These connected components are the parsing analogues to the constituents awaiting expansion in Yngve's analysis.

capturing different components of the variance in reading times. Since integration cost was a significant predictor of reading times for nouns and verbs (i.e. not closed class words) and embedding depth was a significant predictor of reading times for closed class words, integration cost and embedding depth should also be complementary to each other.

5 Methods

5.1 Corpora

We used two corpora in this work. The **English and Simple English Wikipedia** corpus of Hwang et al. (2015, ESEW) is a new corpus of more than 150k sentence pairs designed to address the flaws of the Parallel Wikipedia Corpus of Zhu et al. (2010, PWKP), which was previously dominant in work on text simplification, by using a more sophisticated method of aligning pairs of English and Simple English sentences. We used the section labeled as having 'good' alignments for our work and assumed that, in every sentence pair, the Simple English sentence should be ranked as easier than the English sentence (*rank=1* $<$ *rank=2* in Table 1). This provides a large corpus with noisy labels, as there are likely to be instances where the English and Simple English sentences are not substantially different or the English sentence is the easier one.[8]

For a more controlled corpus, we use Vajjala's (2015) **One Stop English** (OSE) corpus. This corpus consists of 1577 sentence triples, drawn from news stories edited to three difficulty levels: elementary, intermediate, and advanced. Vajjala used $TF * IDF$ and cosine similarity scores to align sentences from stories drawn from `onestopenglish.com`. While One Stop English does not publish an explanation of their methods for creating these texts, they are at least created by human editors for pedagogical purposes, so the labels should be more consistent and reliable than those associated with the ESEW corpus.

The three levels of OSE make it possible to compare system performance on sentence pairs which are close to one another in difficulty (e.g. 'advanced' versus 'intermediate' sentences) with performance on pairs which are further apart, as with 'advanced' sentences paired with their 'elementary' counterparts. In this paper we will refer to the pairs of advanced and elementary sentences as OSE_{far}, the remaining pairs as OSE_{near}, and the full OSE dataset as OSE_{all}. An example triple of sentences from the corpus is given in Table 2.

5.2 Feature Extraction and Feature Sets

We used two parsers to extract 22 features from the corpora. The `ModelBlocks` parser provided features based on surprisal and embedding depth while the Stanford parser[9] provided the dependency parses used to calculate integration cost and idea density features. Both parsers are trained and perform near the state of the art on the standard sections of the Wall Street Journal section of the Penn Treebank.

From `ModelBlocks`' complexity feature extraction mode, we took the lexical and syntactic surprisal features. We used the average lexical surprisal and average syntactic surprisal as idealized measures of the channel capacity required to read a sentence. While this underestimates the channel capacity required to process a sentence, it is at least internally consistent, insofar as a sentence with higher average surprisal overall is likely to require a higher channel capacity as well. We also used the maximum of each form of surprisal as a measure of the maximum demand on cognitive resources. These features comprise the SURPRISAL model.

We also calculated average and maximum values for the embedding depth and embedding difference output from `ModelBlocks`. The average provides an estimate of the typical memory load throughout a sentence, while the (absolute) embedding difference is a measure of how many times a reader needs to push or pop a connected component to or from their memory store. These features comprise the EMBEDDING model.

To extract the remaining features, we first ran the Stanford dependency parser on both corpora. The program `icy-parses` uses part-of-speech tags and head-dependent relations to determine the total, average, and maximum integration cost across a sentence. Here average integration cost functions as another kind of memory load estimate while the maximum value models the most-

[8] Indeed, $37,095$ of the $154,805$ sentence pairs have the same sentence for both English and Simple English Wikipedia and were therefore excluded from our experiments.

[9] `http://nlp.stanford.edu/software/lex-parser.shtml`

Rank	Sentence
2	Gingerbread was brought to Europe in 992 by the Armenian monk Gregory of Nicopolis -LRB- Gregory Makar -RRB- -LRB- Grégoire de Nicopolis -RRB- .
1	Armenian monk Gregory of Nicopolis -LRB- Gregory Makar -RRB- -LRB- Grgoire de Nicopolis -RRB- brought ginger bread to Europe in 992 .

Table 1: Example sentences from English (2) and Simple (1) English Wikipedia.

Rank	Sentence
3	It is a work-hard, play-hard ethic that many of the world's billionaires might subscribe to but it would be a huge change for most workers and their employers.
2	It is a 'work-hard, play-hard' way of thinking that many of the world's billionaires might agree with but it would be a huge change for most workers and their employers.
1	Many of the world's billionaires might agree with this way of thinking but it would be a very big change for most workers and their employers.

Table 2: Example sentences from One Stop English, at levels advanced (3), intermediate (2), and elementary (1). The pair 3–1 is in OSE_{far}, the pairs 3–2 and 2–1 are in OSE_{near}, and all three pairs are in OSE_{all}.

difficult-to-integrate point in the sentence. These features comprise the INTEGRATIONCOST model.

Finally, we use a modified version of the `IDD3` library from Andre Cunha (Cunha et al., 2015) to extract idea density decomposed across three types of propositional idea: predications, modifications, and connections.[10] Here we use only averaged features, as the crucial measure is the idea *density* rather than the raw number of ideas being expressed. These features comprise the IDEADENSITY model.

As a point of comparison for these models, we created a BASELINE which used only sentence length and the average word length as features.

We also created models based on features grouped by the parser used to extract them: SURPRISAL+EMBED for the `ModelBlocks` parser and IDEA+INTEGRATION for the Stanford parser. While `ModelBlocks` achieves competitive accuracies, it is much slower than other state-of-the-art parsers available today. Therefore we wanted to provide a point of comparison regarding the relative utility of these parsers: grouping features by parser allows us to assess the trade-off between model accuracy and the time necessary for feature extraction.

Finally, we considered combinations of the parser-grouped features with the baseline (BASE+SURPRISAL+EMBED and BASE+IDEA+INTEGRATION) and a FULLMODEL using the baseline features and all of the psycholinguistic features.

Replication The scripts required for replication are available at `https://github.com/dmhowcroft/eacl2017-replication`. This includes pointers to the corpora, preprocessing scripts and settings for the parsers, as well as scripts for feature extraction and running the averaged perceptron model.

5.3 Ranking as Classification

In order to rank sentences, we need some way of generating a complexity score for each sentence. Using a perceptron model allows us to train a simple linear scoring model by converting the ranking task into a classification task.

Suppose we have two sentences s_1 and s_2 with feature vectors $\mathbf{s_1}$ and $\mathbf{s_2}$ such that s_1 is more complex than s_2. Then we want to train a perceptron model such that

$$score(s_1) > score(s_2) \quad (2)$$

$$\mathbf{W} \cdot \mathbf{s_1} > \mathbf{W} \cdot \mathbf{s_2} \quad (3)$$

$$\mathbf{W} \cdot (\mathbf{s_1} - \mathbf{s_2}) > 0 \quad (4)$$

We refer to the vector $\mathbf{s_1} - \mathbf{s_2}$ as a vector of *difference features*. In order to train the model, we take all pairs of sentences present in a given corpus and create a difference vector as above. In half of the cases, we flip the sign of the difference vector, creating a binary classification task with balanced classes. The learning problem is now to

[10] Code available at: `https://github.com/dmhowcroft/idd3`.

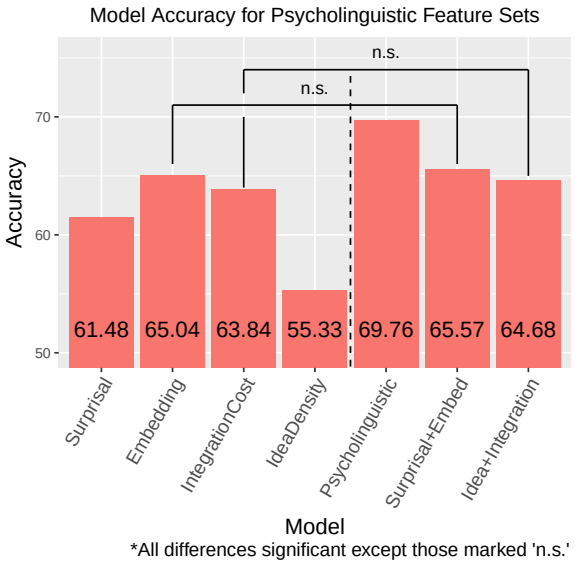

Figure 1: Results on the ESEW corpus for each set of psycholinguistic features individually (first 4 columns) and altogether (5th column), with the feature sets based on the `ModelBlocks` and Stanford parsers in the last two columns.

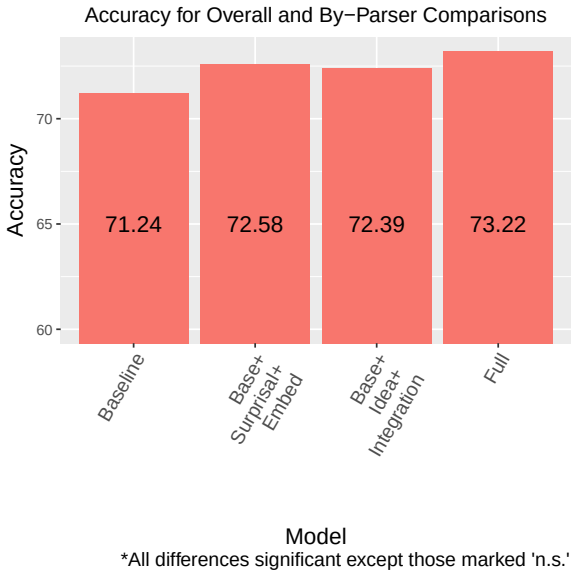

Figure 2: Results for the baseline model, our two parser-grouped feature sets, and the full model on the ESEW corpus.

classify each difference vector based on whether the first term in the difference was the 'easier' or the 'harder' sentence

Note that the benefit to this approach is that the resulting weight vector $\mathbf{W}$ learned via the classification task can be used directly to score individual sentences as well, with the expectation that higher scores will correspond to more difficult sentences.

We use an averaged perceptron model (Collins, 2002) implemented in Python as our classifier.

6 Analysis & Results

The feature sets for individual psycholinguistic theories only achieve accuracies between 55% and 65% (see the first 4 columns of Fig. 1). Combining all of these features into the PSYCHOLINGUISTIC model improves performance to nearly 70% (column 5). Looking at the feature sets grouped by parser (columns 6 and 7), we see that the combination of surprisal and embedding depth (from the `ModelBlocks` parser) significantly outperforms the combination of integration cost and idea density (from the Stanford Parser). However, the strength of the features derived from `ModelBlocks` seems to be primarily driven by the EMBEDDING features, while the strength of the dependency-parse-derived features appears to stem from INTEGRATIONCOST.

Moving to Figure 2, we see that our BASELINE features achieved an accuracy of 71.24%, despite using only average word length and sentence length. This is 1.48 percentage points higher than the 69.76% accuracy of the PSYCHOLINGUISTIC model, which includes surprisal, embedding depth, integration cost, and idea density. However, the FULL model (column 4) outperforms the BASELINE by a statistically significant[11] 1.98 percentage points ($p \ll 0.01$). This confirms our primary hypothesis: psycholinguistic features based on online sentence processing can improve models of sentence complexity beyond a simple baseline.

To address the secondary hypothesis, we turn to the OSE data in Figure 3. The best model for this corpus uses the baseline features combined with embedding depth and surprisal features extracted from `ModelBlocks`. In both OSE_{far} and OSE_{near} we gain about 3 points over the baseline when adding these features (3.18 and 3.25 points, respectively), which is similar to the gains for the FULL model over the baseline. The fact that the increase in performance between the BASELINE model and the best performing model does not differ between the OSE_{near} and the OSE_{far} datasets suggests a lack of support for our secondary hypothesis that these features are espe-

[11] Using McNemar's (1947) test throughout, as is standard for paired samples like ours, with Bonferroni correction where appropriate.

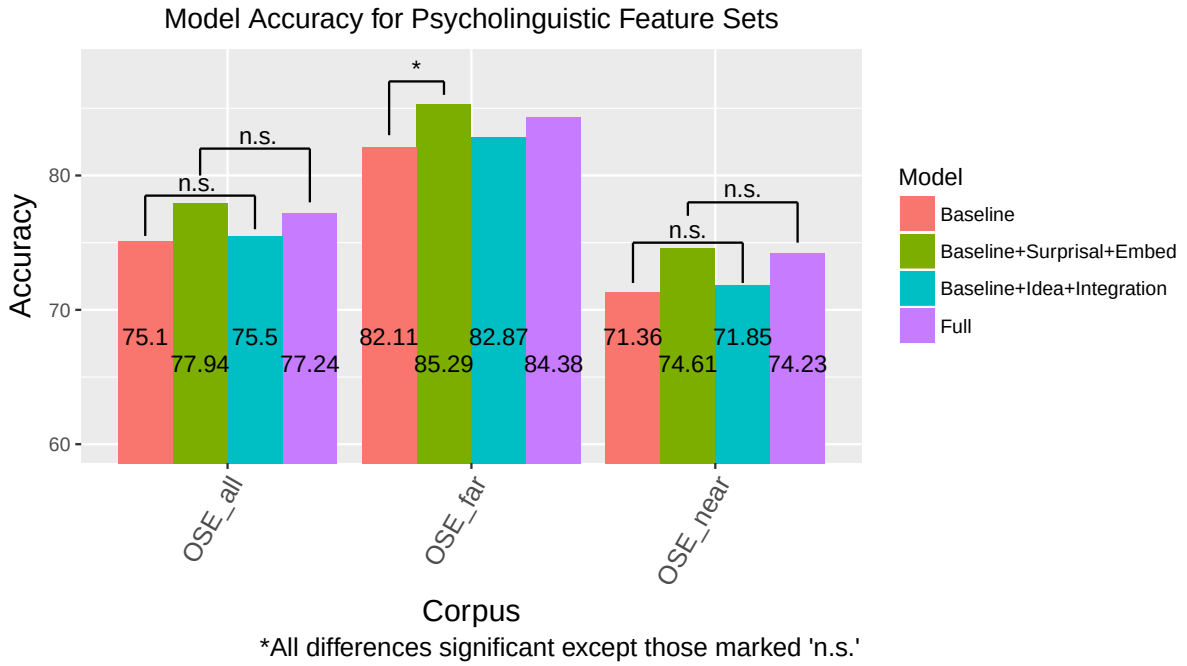

Figure 3: Results for the baseline model, our two parser-grouped feature sets, and the full model on the OSE corpus, with additional breakdown by level proximity.

cially helpful for distinguishing items of similar difficulty levels.

These results warrant a full comparison to the work of Ambati et al. (2016), despite the differences in our evaluation sets. Ambati et al. found that their features based on incremental CCG derivations achieved an accuracy of 72.12%, while the offline psycholinguistic features of Vajjala & Meurers came in at 74.58%, 1.36 percentage points better than our 73.22%. Finally, a model combining all of Vajjala & Meurers featurs with the incremental CCG features achieved a performance of 78.87%. Since the features examined in our study are complementary to those proposed by these two previous studies, a model combining all of these features should further improve in accuracy.

7 Conclusion

We examined features for the ranking of sentences by their complexity, training linear models on two corpora using features derived from psycholinguistic theories of online sentence processing: idea density, surprisal, integration cost, and embedding depth.

Surprisal coupled with embedding depth and our baseline features (average word length & sentence length) performed as well as the full model across all subsets of the OSE corpus. Integration cost and idea density were less effective, suggesting that the gain in speed from running a faster dependency parser may not be worth it. Instead, it is necessary to use the slower `ModelBlocks` parser to extract the more useful features.

Overall, our strongest model combined the baseline features and the online psycholinguistic features. Because these features are complementary to features which have been explored in other work (Vajjala and Meurers, 2014; Ambati et al., 2016), the next step in future work is to combine all of these features and conduct a more comparison between the features proposed here and those examined in earlier work. In the meantime, we have demonstrated that features derived from psycholinguistic theories of sentence processing can be used to improve models for ranking sentences by readability.

Acknowledgments

Thanks are due to Matthew Crocker, Michael White, Eric Fosler-Lussier, William Schuler, Detmar Meurers, Marten van Schijndel, and Sowmya Vajjala for discussions and guidance during the development of this work. We are supported by DFG collaborative research center SFB 1102 'Information Density and Linguistic Encoding'.

References

Ram Bharat Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057. Association for Computational Linguistics.

Lucy Bryant, Elizabeth Spencer, Alison Ferguson, Hugh Craig, Kim Colyvas, and Linda Worrall. 2013. Propositional Idea Density in aphasic discourse. *Aphasiology*, (July):1–18, jun.

Jeanne S. Chall. 1958. *Readability: an appraisal of research and application*. The Ohio State University, Columbus, OH, USA.

Michael Collins. 2002. Ranking Algorithms for NamedEntity Extraction: Boosting and the Voted Perceptron. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 489–496, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Scott A. Crossley, David B. Allen, and Danielle S. McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1):84–101.

Andre Luiz Verucci Da Cunha, Lucilene Bender De Sousa, Leticia Lessa Mansur, and Sandra Maria Aluísio. 2015. Automatic Proposition Extraction from Dependency Trees: Helping Early Prediction of Alzheimer's Disease from Narratives. *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 127–130.

Vera Demberg and Frank Keller. 2008. Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, 109(2):193–210.

William H. Dubay. 2007. *Unlocking Language: The Classic Readability Studies*.

Sarah Tomaszewski Farias, Vineeta Chand, Lisa Bonnici, Kathleen Baynes, Danielle Harvey, Dan Mungas, Christa Simon, and Bruce Reed. 2012. Idea density measured in late life predicts subsequent cognitive trajectories: Implications for the measurement of cognitive reserve. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 67 B(6):677–686.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 229–237.

Alison Ferguson, Elizabeth Spencer, Hugh Craig, and Kim Colyvas. 2014. Propositional Idea Density in women's written language over the lifespan: Computerized analysis. *Cortex*, 55(1):107–121, jun.

M. Galley and K. McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proc. of the 18th International Joint Conference on Artificial Intelligence*, pages 1486–1488.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In Y Miyashita, A Marantz, and W O'Neil, editors, *Image, Language, Brain*, chapter 5, pages 95–126. MIT Press, Cambridge, Massachusetts.

Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34:286–310.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

William S. Gray and Bernice E. Leary. 1935. *What makes a book readable*. University of Chicago Press, Chicago, Illinois, USA.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Denver, Colorado, USA.

T. Florian Jaeger. 2006. *Redundancy and Syntactic Reduction in Spontaneous Speech*. Unpublished dissertation, Stanford University.

T. Florian Jaeger. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62, aug.

David Kauchak, Obay Mouradi, Christopher Pentoney, and Gondy Leroy. 2014. Text simplification tools: Using machine learning to discover features that identify difficult text. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2616–2625.

Alan Kennedy and Joël Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2):153–168.

J. Peter Kincaid, Robert P. Fishburne, Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report, Naval Technical Training Command, Memphis - Millington, TN, USA.

Walter Kintsch and Janice Keenan. 1973. Reading Rate and of Propositions Retention as a Function of the Number in the Base Structure of Sentences. *Cognitive Psychology*, 5:257–274.

Walter Kintsch. 1972. Notes on the structure of semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of memory*, pages 247–308. Academic Press, New York, New York, USA.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems 20 (NIPS)*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–77, mar.

Bertha A. Lively and S. L. Pressey. 1923. A Method for Measuring the 'Vocabulary Burden' of Textbooks. *Educational Administration and Supervision*, IX:389–398.

Iain Macdonald and Advaith Siddharthan, 2016. *Proceedings of the 9th International Natural Language Generation conference*, chapter Summarising News Stories for Children, pages 1–10. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

George A. Miller. 1956a. Human memory and the storage of information. *IRE Transactions on Information Theory (IT-2)*, 2(3):129–137, Sep.

George A. Miller. 1956b. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*.

Courtney Napoles and Mark Dredze. 2010. Learning Simple Wikipedia : A Cogitation in Ascertaining Abecedarian Language. *Computational Linguistics*, (June):42–50.

Charles A. Perfetti. 1969. Lexical density and phrase structure depth as variables in sentence retention. *Journal of Verbal Learning and Verbal Behavior*, 8(6):719–724.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:89–106.

Sarah E. Petersen. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. PhD thesis, University of Washington.

Kathryn P. Riley, David A. Snowdon, Mark F. Desrosiers, and William R. Markesbery. 2005. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. *Neurobiology of Aging*, 26(3):341–347, Mar.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333. Association for Computational Linguistics.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

L. A. Sherman. 1893. *Analytics of Literature*. Ginn & Company, Boston, Massachusetts, USA.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden. Association for Computational Linguistics.

Sowmya Vajjala. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Phd thesis, Eberhard Karls Universitaet Tuebingen.

Marten van Schijndel, Andy Exley, and William Schuler, 2012. *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, chapter Connectionist-Inspired Incremental PCFG Parsing, pages 51–60. Association for Computational Linguistics.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–40.

Michael D. Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–11.

Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. 2010. Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1189–1198. Association for Computational Linguistics.

Victor H. Yngve. 1960. A Model and an Hypothesis for Language Structure. *American Philosophical Society*, 104(5):444–466.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361. Coling 2010 Organizing Committee.