Detecting the “Point of Originality” in Student Writing

Brandon White
Brandeis University
415 South St., Waltham MA, 02454
brandonw@brandeis.edu

Johann Larusson
Brandeis University
415 South St., Waltham MA, 02454
johann@brandeis.edu

ABSTRACT
This paper proposes a new method for the objective evaluation of student work through the identification of original content in writing assignments. Using WordNet as a lexical reference, this process allows instructors to track how key phrases are employed and evolved over the course of a student’s writing, and to automatically determine and visualize the point at which the student’s language first demonstrates original thought, phrased in their own, original words. After sketching the method for isolating “points of originality,” the paper provides a method for visualizing the resultant information. By visualizing otherwise subjective information in a way that is objectively intelligible, the goal provides educators with the ability to monitor student investment in concepts from the course syllabus, and to extend or modify the boundaries of the syllabus in anticipation of pre-existing knowledge or trends in interest.

Author Keywords
Evaluation methods, information visualization and presentation

ACM Classification Keywords
H.5.2: Information interfaces and presentation: user interfaces

INTRODUCTION: SELF-MONITORING IN EDUCATION
For most if not all learning activities, a substantial amount of the instructor’s time and effort is devoted to evaluating and monitoring the quality of students’ work, and thus, hopefully, the depth of their learning [2]. The purpose of this monitoring, however, is not merely the determination of grades; part of the instructor’s work is entirely self-reflective, enabling the instructor to concurrently, or ideally even preemptively, intervene to make adjustments to course pedagogy based on students’ engagement or understanding [9]. While assigning grades might be facile, some difficulties complicate this second objective: how might an instructor intuit when, precisely, students have understood the material sufficiently? Making these determinations manually would prove an intensely laborious and time-consuming process, far more complicated than simple reading and re-reading of any single student’s work.

However, higher education is increasingly requiring students to hand in materials in electronic form, which enables the creation of computer-assisted instructional aids [6]. This paper proposes an automated solution that can be used by educators to help resolve these tensions. Through the application of lexical analysis of student writing, we propose to track how a student’s written language migrates from mere paraphrase to mastery, isolating the moment when the student’s understanding of core concepts best demonstrates the pedagogical principle of recasting, a moment that we’ve chosen to call the “point of originality.” This process recreates the same cognitive activity that educators might ordinarily undergo, yet in an automatic manner that is far less labor intensive. Ultimately, the resulting data is presented to the instructor by way of custom visualizations, which allow the instructor to engage in continuous self-monitoring with minimally expended effort.

ORIGINALITY IN STUDENT WRITING
When students engage in a writing activity, the final evaluation of their work cannot take into account only whether or not the student has provided the most closely correct answer. Process is just as relevant to student writing as content [16]. Student writing that is exceptional is generally seen as that which demonstrates a mastery of the course material in new, profound or statistically unusual ways [11]. The ideal is not only for students to confirm that they’ve understood lectures, but to do so in ways that even the educator might not have thought of.

This process of mastery need not take place all at once. As a student is continually exposed to the same material, or is given the independent opportunity to rethink, reframe, or revisit that material [17], their writing on the subject has the chance to evolve, from rote regurgitation to wholly original expression [12]. At the level of language, this evolution is reflected through recasting.

Recasting is the learning process whereby a student refines his or her understanding of a concept found in course lectures or readings by putting that concept into his or her own words [15]. In the acquisition of new languages especially, this process can be useful, because it allows
students to acquire new vocabulary using the assortment of words already available to them [10, 15]. Even where the student’s understanding of a language is not an explicit concern, recasting can mark a student’s attempts to graduate to more sophisticated or professionalized terminology, or, inversely but to the same end, to place new concepts into terms that are nearer to what the student would naturally be more likely to say [3]. This process of learning aligns with theories of schema formulation [7], as well as the express principles of educational constructivism [8]. Attempts at recasting in order to transition from one supplied concept to a related one thus bears further resemblance to the sensemaking process known as “scaffolding” [5].

For an instructor, the simple identification of recast terminology within a student’s written work can provide an effective barometer for pedagogical self-reflection. If a subset of terms or concepts are deemed vital to the syllabus, repetitions and recast iterations of those same terms will at least suggest that those terms are being acknowledged and reflected upon. Yet if the instructor hopes not only to identify instances where key concepts are deployed, but to determine how comprehensively the concepts are being internalized, it is first necessary to possess a method of scoring how original any given recast might be. In order to do this, we propose a metric for isolating a specific point of originality within student writing.

COMPUTER-ASSISTED DETECTION

The continuing migration of more and more teaching materials to digital venues [6], and thus a machine-readable form, has the supplementary benefit of making a student’s day-to-day learning activities more transparent in each stage of the teaching process. Through computer mediation, this large body of written work can be compared against the linguistic data present within a database such as WordNet.

WordNet is a lexical database that arranges nouns, verbs, adjectives, and adverbs by their conceptual-semantic and lexical relationships [4]. Whereas a simple thesaurus would be able to identify any two words as synonyms or antonyms for the same concept, WordNet is able to note the similarity between two words that don’t have literally identical meanings. These relationships are ideally meant to mirror the same lexical associations made by human cognition.

WordNet’s arrangement is hierarchical, which is to say that certain terms are more closely related than others. Although a dog and an alligator are both animals, a dog ought to be more closely related to a cat by virtue of the fact that alligators are not domestic animals. Within WordNet, these relationships are displayed as “synsets,” clusters of terms that branch, like neurons or tree branches, from more specific to more and more diffuse associations (see Figure 1). If two words are found within one another’s synset tree, it stands to reason that these terms are, in some way, related, be it closely, as between a dog and a cat, or distantly, as between a dog and a schoolhouse, which are related only insofar as both are entities, or nouns. As discussed in the next section, these distances between two terms can be calculated, and assigned a value commensurate with their degree of semantic relatedness [1].

If an instructor wanted to know how frequently the students within a class had made reference to a key term, say, in a perhaps trivial example, “democracy,” then the instructor might simply search the electronic text for direct recurrences of those words. Yet it is possible, indeed, likely, that students might write about their understanding of democracy without making direct reference to the phrase itself: students might just as easily substitute synonymous phrases, like “republicanism,” or related but not directly analogous terms, like “voting,” “representation,” or “authoritarianism.” To determine that any word is a likely recast of another, it is not enough, however, merely to determine that the two terms are related. Only by determining how, or to what extent, two terms are related is it possible to judge the degree to which original recasting is taking place.

---

![Model synset tree](image)

**Figure 1. Model synset tree (by hyponym relation)**

**DETERMINING THE RELATIONS BETWEEN WORDS**

The hierarchical arrangement inherent to WordNet provides one method of determining the relationship between two terms. If the synset tree of one term encompasses another term, it is simple enough to note how many synset jumps it takes to move from one to another. In Figure 1, a “Dalmatian” is a type of “dog,” which itself belongs to the subcategory of “domestic animals;” thus there are two tiers of associations between the concepts of “Dalmation” and “domestic animals.” Unfortunately, however, just how closely any two terms might be related is not a straightforward, linear relationship. WordNet organizes related terms by their precise lexical entailment, such that nouns might be categorized as synonyms, hypernyms, hyponyms, holonyms and meronyms, and verbs might be categorized as hypernyms, troponyms, or entailment terms as seen in Table 1.

These possible entailments provide a rudimentary roadmap for all the ways in which two words might be related. Since WordNet attempts to map the cognitive associations automatically formed between words [4], a student’s evocation [13] of the holonym or hypernym of a given noun
Let $W$ be a set of all words from a given student assignment and let $S$ be a set of stopwords to be omitted. Then, $M$, the set of synset terms found in any given writing sample can be defined as:

$$M = Q \cap (W - S)$$

WordNet stores synset matches in a tree structure with $\beta$ as the root node. Then, $\delta$, the distance (depth) for any given synset term in $M$ from the root node (query term) $\beta$ is defined as:

$$\delta = \begin{cases} 
0 & \text{if } \alpha = \beta \\
1 & \text{if } \alpha \text{ is first child of } \beta \\
2 & \text{if } \alpha \text{ is second child of } \beta \\
\vdots & 
\end{cases}$$

WordNet also supplies the type of each synset term. Thus, $t$, the “word type” of any given synset term in $M$, is defined as:

$$t = \begin{cases} 
1.0 & \text{if } \alpha = \text{synonym/antonym} \\
0.9 & \text{if } \alpha = \text{holonym/meronym} \\
0.85 & \text{if } \alpha = \text{synonym} \\
0.7 & \text{if } \alpha = \text{hypernym/hyponym} \\
\end{cases}$$

Then, the weight for any given synset term in $M$ in the general “point of originality” estimate is calculated as follows:

$$\alpha = (\delta \times 0.7) \times t$$

$\delta$, the depth for any given synset term, is multiplied by a constant value of 0.7, which reflects the diminished associations between terms the farther separated two terms are along the synset tree. This value is selected because it corresponds with the calculation of distance between terms that yields the nearest match with human intuition [18].

Then, $P$, the point of originality in a given student’s writing for the query term $\beta$, can be defined as:

$$P(\beta) = \sum_{n=0}^{\frac{|M|}{\alpha_n}}$$

This same calculation is then performed for all of the writing samples by a given student. Once the point of originality is calculated, all instances of originality are plotted upon a horizontal timeline for optimal instructor comprehension.

### CALCULATING THE POINT OF ORIGINALITY

The visualization for determining the point of originality in student writing depends upon the input of a specific query term by the instructor, after which, a distance calculation is performed. Query terms need to be input manually only because this affirms the pedagogical utility of the process; inputting a key syllabus term allows the instructor to see recasts of that term across the entire body of student writing in a course.

The process then generates the WordNet synset trees for each of the input query terms. When the synset trees of the preserved input terms are retrieved, words within those trees are compared. Where matches are found, a distance calculation between the original query term and the match within the student’s work is performed. This calculation takes the form of:

Let $\beta$ be a query term supplied by the instructor. Then, let $Q$ be a set containing all synset word matches from the WordNet database for $\beta$. Yet while this simple index displays just how any two terms might be related, all the possible relationships noted are not necessarily equal. Some relationships, like that between synonyms *smile* and *grin*, are obviously bound to be more strongly associated than that between *mammal* and *dog*. Following a method first noted by Yang & Powers [18], it is possible to install a series of weights that can best calculate the semantic distance between any two terms. This method, in particular, is useful because of all possible methods, it bears the highest correspondence between its own distance calculations and the intuitions of actual human respondents (at 92.1 percent accuracy). The application of this distance calculation is explained in greater detail in the following section.

### Table 1. Possible lexical entailments in WordNet

<table>
<thead>
<tr>
<th>Noun</th>
<th>Synonym: X is a synonym of Y if X means Y</th>
<th>Example: [smile, grin]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym: X is a hypernym of Y if every X is a kind of Y</td>
<td>Example: [dog, mammal]</td>
<td></td>
</tr>
<tr>
<td>Hyponym: X is a hyponym of Y if every Y is a kind of X</td>
<td>Example: [mammal, dog]</td>
<td></td>
</tr>
<tr>
<td>Holonym: X is a holonym of Y if Y is part of X</td>
<td>Example: [hand, finger]</td>
<td></td>
</tr>
<tr>
<td>Meronym: X is a meronym of Y if X is part of Y</td>
<td>Example: [finger, hand]</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>Synonym: X is a synonym of Y if X means Y</td>
<td>Example: [jump, leap]</td>
</tr>
<tr>
<td>Hypernym: X is a hypernym of Y if every X is a kind of Y</td>
<td>Example: [drive, travel]</td>
<td></td>
</tr>
<tr>
<td>Troponym: X is a troponym of Y if X is likewise doing Y</td>
<td>Example: [lisp, speak]</td>
<td></td>
</tr>
<tr>
<td>Entailment: X entails Y if Y requires doing X</td>
<td>Example: [sleeping, snoring]</td>
<td></td>
</tr>
</tbody>
</table>

**CALCULATING THE POINT OF ORIGINALITY**

The visualization for determining the point of originality in student writing depends upon the input of a specific query term by the instructor, after which, a distance calculation is performed. Query terms need to be input manually only because this affirms the pedagogical utility of the process; inputting a key syllabus term allows the instructor to see recasts of that term across the entire body of student writing in a course.

The process then generates the WordNet synset trees for each of the input query terms. When the synset trees of the preserved input terms are retrieved, words within those trees are compared. Where matches are found, a distance calculation between the original query term and the match within the student’s work is performed. This calculation takes the form of:

Let $\beta$ be a query term supplied by the instructor. Then, let $Q$ be a set containing all synset word matches from the WordNet database for $\beta$. Yet while this simple index displays just how any two terms might be related, all the possible relationships noted are not necessarily equal. Some relationships, like that between synonyms *smile* and *grin*, are obviously bound to be more strongly associated than that between *mammal* and *dog*. Following a method first noted by Yang & Powers [18], it is possible to install a series of weights that can best calculate the semantic distance between any two terms. This method, in particular, is useful because of all possible methods, it bears the highest correspondence between its own distance calculations and the intuitions of actual human respondents (at 92.1 percent accuracy). The application of this distance calculation is explained in greater detail in the following section.

**Table 1. Possible lexical entailments in WordNet**

<table>
<thead>
<tr>
<th>Noun</th>
<th>Synonym: X is a synonym of Y if X means Y</th>
<th>Example: [smile, grin]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym: X is a hypernym of Y if every X is a kind of Y</td>
<td>Example: [dog, mammal]</td>
<td></td>
</tr>
<tr>
<td>Hyponym: X is a hyponym of Y if every Y is a kind of X</td>
<td>Example: [mammal, dog]</td>
<td></td>
</tr>
<tr>
<td>Holonym: X is a holonym of Y if Y is part of X</td>
<td>Example: [hand, finger]</td>
<td></td>
</tr>
<tr>
<td>Meronym: X is a meronym of Y if X is part of Y</td>
<td>Example: [finger, hand]</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>Synonym: X is a synonym of Y if X means Y</td>
<td>Example: [jump, leap]</td>
</tr>
<tr>
<td>Hypernym: X is a hypernym of Y if every X is a kind of Y</td>
<td>Example: [drive, travel]</td>
<td></td>
</tr>
<tr>
<td>Troponym: X is a troponym of Y if X is likewise doing Y</td>
<td>Example: [lisp, speak]</td>
<td></td>
</tr>
<tr>
<td>Entailment: X entails Y if Y requires doing X</td>
<td>Example: [sleeping, snoring]</td>
<td></td>
</tr>
</tbody>
</table>

**CALCULATING THE POINT OF ORIGINALITY**

The visualization for determining the point of originality in student writing depends upon the input of a specific query term by the instructor, after which, a distance calculation is performed. Query terms need to be input manually only because this affirms the pedagogical utility of the process; inputting a key syllabus term allows the instructor to see recasts of that term across the entire body of student writing in a course.

The process then generates the WordNet synset trees for each of the input query terms. When the synset trees of the preserved input terms are retrieved, words within those trees are compared. Where matches are found, a distance calculation between the original query term and the match within the student’s work is performed. This calculation takes the form of:

Let $\beta$ be a query term supplied by the instructor. Then, let $Q$ be a set containing all synset word matches from the WordNet database for $\beta$. Yet while this simple index displays just how any two terms might be related, all the possible relationships noted are not necessarily equal. Some relationships, like that between synonyms *smile* and *grin*, are obviously bound to be more strongly associated than that between *mammal* and *dog*. Following a method first noted by Yang & Powers [18], it is possible to install a series of weights that can best calculate the semantic distance between any two terms. This method, in particular, is useful because of all possible methods, it bears the highest correspondence between its own distance calculations and the intuitions of actual human respondents (at 92.1 percent accuracy). The application of this distance calculation is explained in greater detail in the following section.

**Table 1. Possible lexical entailments in WordNet**

<table>
<thead>
<tr>
<th>Noun</th>
<th>Synonym: X is a synonym of Y if X means Y</th>
<th>Example: [smile, grin]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym: X is a hypernym of Y if every X is a kind of Y</td>
<td>Example: [dog, mammal]</td>
<td></td>
</tr>
<tr>
<td>Hyponym: X is a hyponym of Y if every Y is a kind of X</td>
<td>Example: [mammal, dog]</td>
<td></td>
</tr>
<tr>
<td>Holonym: X is a holonym of Y if Y is part of X</td>
<td>Example: [hand, finger]</td>
<td></td>
</tr>
<tr>
<td>Meronym: X is a meronym of Y if X is part of Y</td>
<td>Example: [finger, hand]</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>Synonym: X is a synonym of Y if X means Y</td>
<td>Example: [jump, leap]</td>
</tr>
<tr>
<td>Hypernym: X is a hypernym of Y if every X is a kind of Y</td>
<td>Example: [drive, travel]</td>
<td></td>
</tr>
<tr>
<td>Troponym: X is a troponym of Y if X is likewise doing Y</td>
<td>Example: [lisp, speak]</td>
<td></td>
</tr>
<tr>
<td>Entailment: X entails Y if Y requires doing X</td>
<td>Example: [sleeping, snoring]</td>
<td></td>
</tr>
</tbody>
</table>
2. This drop-down menu allows the instructor to select which student’s writing samples are currently being displayed.

3. This timeline displays the date/times of each of the student’s writing samples. By default, all markers share the same default color, a monochrome black. Writing samples are then color-coded differently, however, depending on the value of the point of originality (P) score for any given writing sample. These colors progress from colder to warmer colors on the ROYGBIV spectrum the higher the value of P. These color assignments present an intuitive way for the instructor to quickly recognize that the sample has been assigned a higher originality value.

4. This window displays the student’s writing samples in excerpted form, with the matches between the query terms and the synset terms found within the writing sample (M) highlighted in the same color as that sample’s marker color. The colored marker itself is displayed as the sample’s bullet point. By default, this window is pre-sorted, from highest P value to lowest.

5. If a writing sample marker is selected, either in the timeline window (see inset 3) or in the list (see inset 4), the text of that writing sample is displayed here, again, with the matched synset terms (M) highlighted.

This assortment of visualization options allows the point of originality calculation to be displayed in a number of intuitive ways: as a list (inset 4), within chronology (inset 3), and in context (inset 5).

This project is now in its implementation phases. Having already gained faculty support, once a working prototype has been developed, the visualization will continue to be refined with actual instructor input through a participatory design framework [14]. In order to validate the actual “point of originality” calculations, the final visualization mechanism will then be applied to a corpus of student writing, in this case three years’ worth of student blog posts on a variety of topics. Human readers (faculty members belonging to the disciplines appropriate to the blog posts’ content) will then read through the same writing samples and make their own suppositions as to when originality in the students’ work has been achieved. This expert intuition will then be compared to the calculated “points of originality,” in order to affirm that the mechanism achieves its intended purpose, and corresponds not only with human intuition, but to the pedagogical judgments of educators.

REFERENCES

