Automatic analysis of thematic structure in written English

Kwanghyun Park and Xiaofei Lu
Myongji University / The Pennsylvania State University

This paper proposes and describes a computational system for the automatic analysis of thematic structure, as defined in Systemic Functional Linguistics, in written English. The system takes an English text as input and produces as output an analysis of the thematic structure of each sentence in the text. The system is evaluated using data from The Wall Street Journal section of the Penn Treebank (Marcus et al. 1993) and the British Academic Written English corpus (Gardner & Nesi 2013). An experiment using these data shows that the system achieves a high degree of reliability in regard to both identifying theme-rheme boundaries and determining several of the linguistic properties of the identified themes, including syntactic nodes, theme function, markedness, mood types, and theme roles. To illustrate how the system is used, we describe an example application designed to compare collections of novice and expert academic writing in terms of thematic structure.

Keywords: automatic theme identification, English writing, Systemic Functional Linguistics, thematic structure

1. Introduction

In Systemic Functional Linguistics (SFL), the ‘theme’ is the element that serves as “the starting-point for the message” and conveys “what the clause is going to be about” (Halliday & Matthiessen 2004:64). It comes at the beginning of the clause and typically carries familiar or given information. The ‘rheme’ is “the part of the clause in which the theme is developed” (Egginis 2004:300). It comes after the theme and typically carries unfamiliar or new information. Together, the theme and rheme constitute the ‘theme system’, and the theme-rheme structure within the clause is referred to as “thematic structure”. The theme system is instrumental to packaging the clause into a message in a way that expresses the intended
meaning (Eggins 2004). Typically, the theme establishes the common ground between the previous message and the following new message whereas the rheme develops ideas and arguments in the new message based on this common ground.

Two important aspects of thematic structure are ‘thematic selection’ and ‘thematic progression’. Thematic selection, that is, what speakers and writers place in the theme, plays an important role in determining the communicative effect of the message. By packaging parts of the message as prioritized or marginal and as given or new, the thematic selection sets up an environment in which the reader can understand and interpret the message (Eggins 2004). Thematic progression, that is, the patterning of thematic choice in a text, makes a critical contribution to ensuring the cohesive development of the text (Eggins 2004, McCabe 1999). For example, the pattern can be a parallel progression where the same theme is used in multiple clauses, or a linear progression where the rheme in the previous clause serves as the theme of the next clause (in a zig-zag pattern). A cohesive text entails the use of well-chosen themes as well as an effective pattern of thematic progression.

Given their central role in the creation of a text, thematic selection and progression are the focus of numerous studies that engage the SFL analysis of language in both academic and non-academic texts. Some of these studies focus on a specific group of writers or a particular type of writing, such as history of science essays by undergraduate students (North 2005), compositions by college-level Chinese learners of English (Wang 2007), academic article abstracts (Ghadessy 1999), and news reports (Gomez 1994). Other studies examine the similarities and differences between different groups of writers, different types of writing, or different components of the same type of writing. Examples of this strand of research include comparative studies of different rhetorical sections of research articles (Martinez 2003), original vs. translated academic texts (Jalilifar 2009), and news reports on the same issue from different regions (Lu 2002), as well as the work of native vs. non-native writers, expert vs. novice writers (Gosden 1995), and international academic vs. local academic writers (Jalilifar 2010). Collectively, these studies have yielded a range of substantial findings in respect to the role of thematic structure in academic and non-academic writing; the relationship of thematic structure to such factors as register, genre, rhetorical structure, and ideology; and the gap between non-native/novice writers and native/professional writers, among others.

Just as important as thematic structure analysis for the SFL analysis of language in language-related research is the annotation of text corpora in regard to thematic structure (and other functional information). In fact, this kind of annotation has been shown to be useful for a number of natural language processing tasks, such as machine translation, information extraction, and information
retrieval (e.g. Kappagoda 2009, Steinberger & Bennett 1994). The manual annotation of thematic structure in large-scale text corpora is a formidable task, however, there is a clear need for computational tools that can automate thematic structure analysis to a high level of accuracy.

Whereas SFL has been incorporated into computer-aided or automated text analysis in various ways in the last two decades, few attempts have been made to fully automate thematic structure analysis for written English. For example, Souter (1996) develops a parser using a transcribed spoken corpus that was annotated using systemic functional grammar. Although the parser analyzes both the units and elements of the sentence structure, it does not identify themes. O’Halloran (2003) develops a software package, Systemics 1.0, for the discourse analysis of text including thematic structure analysis within the SFL framework. This tool is very useful for facilitating computer-aided analysis of text; however, it does not have any automatic text analysis capabilities. Kappagoda (2009) discusses the rationale for using SFL in automated text mining, develops a grammatical annotation scheme to enrich text corpora, and trains a machine learner to automatically annotate word functions in the corpora. Thematic structure is discussed in only a minimal way in this work, however, and theme identification does not feature in the annotation scheme.

The most relevant effort on this front is probably Schwarz et al.’s (2008) rule-based system for automatic theme identification. In this system, a set of rules is manually derived from a collection of sentences that have all been syntactically parsed by the Stanford parser (Klein & Manning 2003) and manually analyzed for theme. A simple pattern-matching algorithm then matches the theme identification rules against the syntactic parses generated for an input text. Using 700 sentences selected from 209 academic abstracts, the authors report sentence coverage (sentences processed) of 89% and overall precision (themes accurately identified) of 81.74%. However, this system is capable of identifying only simple themes, i.e. the ones with a single thematic element, not multiple themes, i.e. the ones with two or more thematic elements (see Section 2 below for a discussion of the distinction between the two).

This paper describes a computational system for the automatic analysis of thematic structure, as defined in SFL, in written English. The system takes an English text in plain text format as input and produces as output an analysis of the thematic structure of each sentence in the text. The system is evaluated using data from a corpus of writings by expert writers, i.e. The Wall Street Journal (WSJ) section of the Penn Treebank1 (Marcus et al. 1993) and a corpus of student writing.

---

i.e. the British Academic Written English (BAWE) corpus\(^2\) (Gardner & Nesi 2013). The results reported in the present paper show that the system is very reliable for identifying theme boundaries and for determining several linguistic properties of the identified themes, including syntactic nodes (i.e. phrasal or clausal categories), theme function, markedness, mood types, and theme roles. To illustrate how the system is used, we describe an example application designed to compare novice and expert academic writing in terms of thematic structure.

The rest of the paper is organized as follows. Section 2 reviews how thematic structure is defined in the SFL literature and summarizes the definitions adopted in the current study. Section 3 describes the design and technical details of the computational system. Section 4 reports the system’s performance on unseen test data from the WSJ and BAWE corpora and presents the results of the error analysis. Section 5 describes how the system is used in an example application to compare thematic structure in novice second language writing from the BAWE corpus and expert academic writing data, collected from an open access academic journal (Language Learning & Technology). Section 6 concludes the paper with a discussion on potential applications of the system and avenues for future research.

2. Theme analysis

In this section, we first discuss the definition of theme as conceptualized in the SFL literature. We then examine the distinctions made in the literature between marked and unmarked themes based on their markedness status as well as those between ‘topical’, ‘interpersonal’, and ‘textual’ themes based on the types of meaning they construe.

2.1 Definition of theme in SFL

Within the SFL framework, the unit of text analysis is the clause, and in terms of thematic structure, each clause can be divided into two functional components that represent the distribution of information in the clause, namely, the theme and the rheme. Structurally, the theme is the first phrase(s) of the clause and provides information that has already been established in the text. As such, the theme serves as “the point of the departure of the message” in the clause and “locates and orients the clause within its context” (Halliday & Matthiessen 2004:64). The

\(^2\) The British Academic Written English (BAWE) corpus was developed at the University of Warwick, Reading University, and Oxford Brookes University. The corpus is accessible at www.coventry.ac.uk/BAWE (accessed October 2014).
rHEME is the rest of the clause after the theme where the information or idea initiated in the theme unfolds and develops. Figure 1, adapted from Thompson (2004), provides two illustrative examples (the themes are underscored).

<table>
<thead>
<tr>
<th></th>
<th>THEME</th>
<th>RHEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><em>For centuries</em> yellow canaries have been used to “test” the air in mining.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td><em>In mining</em> yellow canaries have been used to “test” the air for centuries.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Theme and rheme

Sentence A is the opening sentence of a news article about an industrial history exhibition. Here, the writer takes a historical perspective (*For centuries*), which is appropriate for the topic of an article about history. Sentence B, derived by reordering Sentence A, has a different theme. Compared with the theme in A, the theme in B (*In mining*) is more restrictive than the theme in Sentence A (Thompson 2004: 143). The choice between potential themes that would each render the same literal meaning is referred to as thematic choice.

In order to create a cohesive and effective text, e.g. a well-developed argumentative essay, therefore, a writer must make choices regarding how to organize the theme and rheme in each clause. Experienced writers typically start with given or known information in the theme position of the clause and then present new information or develop an argument in the remaining part of the clause.

### 2.2 Marked and unmarked theme

A theme can have one of two kinds of status, marked or unmarked, depending on the mood class of the clause (i.e. declarative, imperative, or interrogative) and whether the theme “conflates with the Mood structure constituent that typically occurs in first position in clauses of that Mood class” (Eggins 2004: 318). For example, the role of the theme in a declarative clause is typically that of Subject, as in *You can’t store protein* where *You* is both the Subject and theme of the clause (the theme is underscored). As a declarative sentence typically begins with a Subject, the choice of theme in this case is ‘unmarked’. Note, however, that the word order in the clause can be changed (e.g. *Protein, you can’t store*) without changing the meaning of the sentence significantly. The choice of theme this time is ‘marked’, as it is less typical to begin a declarative clause with an Object. The same principle can be applied to questions and commands. That is, a question typically begins with a verb (e.g. *Can you help?*) or a WH element (e.g. *How can I help you?*), whereas a command or a request typically begins with a verb that represents the desired action (*Help me!*). Table 1 shows the unmarked choices in the theme system.
Table 1. Unmarked themes

<table>
<thead>
<tr>
<th>Mood class</th>
<th>Unmarked theme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declarative</td>
<td>Subject</td>
<td><em>She wants to read a book.</em></td>
</tr>
<tr>
<td>Imperative</td>
<td>Predicicator</td>
<td><em>Give her the book.</em></td>
</tr>
<tr>
<td>Polar interrogative</td>
<td>Finite</td>
<td><em>Did you give her the book?</em></td>
</tr>
<tr>
<td>WH-interrogative</td>
<td>WH element</td>
<td><em>Why didn’t you give her the book?</em></td>
</tr>
</tbody>
</table>

When the choice of theme does not conform to the unmarked thematic scheme in Table 1, the theme is marked. Unmarked themes represent the default or usual choice in language use, whereas marked themes generally signal some kind of special effect or “that something in the context requires an atypical meaning to be made” (Eggins 2004: 318).

2.3 Multiple themes

If a theme incorporates more than one type of thematic element, it is referred to as a multiple theme. The types of thematic elements correspond to the three types of meaning expressed in the theme, which are referred to as ‘topical’, ‘interpersonal’, or ‘textual’. A topical theme expresses the content, that is, *what* is being expressed; a textual theme helps create cohesion in the text; and an interpersonal theme expresses the speaker’s attitude toward the interlocutor or the experience being described. Structurally, the theme comprises one and only one topical element, i.e. a constituent that is a participant, a circumstance, or a process. In addition, a textual or interpersonal theme can precede the topical theme, giving rise to a multiple theme, as illustrated in Figure 2.

```
And               oddly,       he           was right
  textual         interpersonal  topical

THEME             RHEME
```

Figure 2. Multiple themes

In Figure 2, the textual theme (*And*) creates cohesion with the prior clause, the interpersonal theme (*oddly*) represents the speaker’s attitude toward the described experience, and the topical theme (*he*) corresponds to a person, the Subject of the message.

It should be noted that theme identification is not always straightforward, as SFL researchers differ in regard to how they define some types of themes. For example, Halliday & Matthiessen (2004:73) consider the existential *there* as a theme in and of itself, whereas Thompson (2004) suggests that the topical element following the existential *there* is also part of the theme. According to Thompson
(2004), *there* serves formally as the Subject of the sentence, but semantically, it does not represent any topical element or express experiential meaning. For this reason, *there* cannot constitute a theme. Therefore, the verb expressing the existential process should be included in the theme (Thompson 2004: 161). Figure 3 shows the discrepancy.

<table>
<thead>
<tr>
<th>There</th>
<th>were three jovial Welshmen</th>
<th>Halliday &amp; Matthiessen (2004)</th>
</tr>
</thead>
<tbody>
<tr>
<td>THEME</td>
<td>RHEME</td>
<td></td>
</tr>
<tr>
<td>There were</td>
<td>three jovial Welshmen</td>
<td>Thompson (2004)</td>
</tr>
<tr>
<td>THEME</td>
<td>RHEME</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3.** Two possible analyses of *there* as theme

Such discrepancies in regard to definitions used by SFL theoreticians present a hurdle for the current study. As the SFL research community tackles these discrepancies, we defaulted to the theme definition that entails the simpler of the two structures represented in Figure 3 — here, Halliday & Matthiessen’s (2004) analysis. This definition had the advantage of rendering the development of our system, i.e. the thematic structure analyzer, less computationally expensive than it would have been had we used Thompson’s (2004) definition instead.

3. System description

The system we describe in this section is the *Theme Analyzer*, a computer program that identifies the thematic structure of English sentences based on syntactic parsing and predefined rules. The *Theme Analyzer* takes an English text as input, and as output returns information about the theme of each sentence in the text. This output includes information about the theme’s boundary (i.e. where the theme begins and ends), syntactic nodes (e.g. NP, PP, or VP), theme function (topical, textual, or interpersonal), markedness status (marked or unmarked), mood type (e.g. declarative, interrogative, or imperative), and theme role (e.g. Subject, Complement, and Adjunct). The thematic analysis is performed via a three-step procedure:

i. Syntactic parsing: preprocessing the input text using a third-party syntactic parser, which returns a parse tree for each sentence in the text that represents an analysis of its syntactic structure;

ii. Theme boundary detection: decomposing the parse trees and determining the boundaries of the themes in the sentences using a set of patterns and rules;

iii. Theme information extraction: obtaining information on each theme, including syntactic node, theme function, markedness, mood type, and theme role.
3.1 Preprocessing

Determining the thematic structure of a clause requires an analysis of its syntactic structure. The Theme Analyzer obtains this information by first parsing the input text with the Stanford parser (Klein & Manning 2003). The parser has built-in sentence segmentation, word segmentation, and part-of-speech (POS) tagging functionalities. In other words, it can identify the boundary of each sentence in the input text, segment each sentence into individual words and punctuation marks, and annotate each of these with a tag indicating its POS category. Finally, the parser analyzes the syntactic structure of each sentence and displays the result of the analysis in a parse tree. The parser follows the conventions of the Penn Treebank (Marcus et al. 1993) for annotating the POS, phrasal, and clausal categories. For instance, for Example (1) from Halliday (1994: 38), the parser returns the parse tree in Example (2)3:

(1) The duke has given my aunt that teapot.

(2) (ROOT
    (S
     (NP (DT The) (NN duke))
     (VP (VBZ has))
     (VP (VBN given))
     (NP (PRP$ my) (NN aunt))
     (NP (DT that) (NN teapot))))

3.2 Theme boundary detection

The Theme Analyzer takes the parse tree generated by the parser and decomposes it into three key structural units: T-units, clauses, and phrases. Each of the T-units is analyzed further to determine whether it has a dependent or an independent clause or is a complex T-unit, as these types are treated differently. Operational definitions of these units are as follows:

i. A ‘T-unit’ refers to one main clause plus any subordinate or nonclausal structure embedded within or attached to it (Hunt 1970: 4). Structurally, then, a T-unit consists of one or more clauses, each of which consists of one or more phrases.

3. In Example (2), ROOT = root of the parse tree, S = simple declarative clause, NP = noun phrase, DT = determiner, NN = singular or mass noun, VP = verb phrase, VBZ = verb (third-person singular present), VBN = verb (past participle), PRP$ = possessive pronoun.
ii. A ‘phrase’ is a group of words under one head constituent that determines the linguistic category of the phrase. The parser tags phrasal nodes with the P suffix. For example, noun phrases are tagged as NP, adjectival phrases as ADJP, verb phrases as VP, and WH-noun phrases as WHNP.

iii. A ‘clause’ is defined as a structure with a subject and a finite verb (Hunt 1965), including independent clauses, nominal clauses, adjectival clauses, and adverbial clauses. The Theme Analyzer identifies clauses and their categories by examining the clausal and phrasal nodes and their relationships in the parse tree. For example, when the analyzer sees a tag that represents a clausal node (i.e. S, SINV, or SQ), it responds by looking for a tag for a finite verb, VBD, VBP, or VBZ.

Once these basic units have been identified, the analyzer activates the set of rules in order to identify the theme. Figure 4 shows the workflow of the Theme Analyzer.

Figure 4 represents the decision-making process of Theme Analyzer. The analyzer starts by decomposing the parse tree to identify T-units. If there is more than one T-unit, the analyzer separates the T-units into single T-units and processes them one by one. If a T-unit contains only one main clause, the analyzer proceeds to find the first topical element in the clause and considers everything from the beginning of the clause up to and including the first topical element as the theme. If there is no topical element serving as a theme (e.g. Fire, fire!), the analyzer evaluates the input as a T-unit with no theme. If a T-unit consists of one main clause and one or more dependent clauses, identification of the theme depends on the position of the dependent clause in the complex T-unit. For example, if the dependent clause precedes the main clause (e.g. If you are late, call me), the entire dependent clause (underscored) is considered the theme. If the main clause precedes the dependent clause (e.g. Call me if you are late), the head constituent of the main clause is the theme.

In the main clause, the theme is the first head phrase (NP, VP, ADVP, PP, or WHNP) that is immediately dominated by the highest clausal node (S, SINV, or SQ). To illustrate this point, Figure 5 shows the clausal and thematic structure of the sentence The duke has given my aunt that teapot. Given this sentence, the analyzer goes down the parse tree until it identifies the T-unit and the main clause. There is no dependent clause. Within the main clause, there are two head nodes representing two phrases, NP and VP. The NP (The duke) is the first head phrase immediately dominated by the clausal node S; therefore, the analyzer identifies

---

4. SINV = inverted declarative sentence, SQ = inverted yes/no question or main clause of a WH-question.
Kwanghyun Park and Xiaofei Lu

the NP as the theme. Figure 5 illustrates the phrasal structure and corresponding thematic structure of the sentence:

<table>
<thead>
<tr>
<th>T-Unit</th>
<th>Main Clause</th>
<th>NP</th>
<th>VP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>The duke</td>
<td>has given my aunt that teapot.</td>
</tr>
<tr>
<td>THEME</td>
<td>RHEME</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Structural definition of theme

3.3 Theme information extraction

When the theme has been identified, the analyzer first establishes the mood class of the clause (e.g. declarative, interrogative, or imperative) and the corresponding role of the thematic element (e.g. Subject or Predicator). Then, the analyzer
determines whether the theme is the typical choice for the mood, or whether the theme “conflates with the Mood structure constituent that typically occurs in the first position in clauses of that Mood class” (Eggins 2004: 318). A list of patterns is adapted from Schwarz et al.’s (2008) rule-based system to obtain the following information about the identified themes: syntactic node (e.g. NP, VP, or ADVP), theme function (topical, interpersonal, or textual), markedness (marked or un-marked), mood type (declarative, interrogative, or imperative), and theme role (e.g. Subject or Complement). The patterns are summarized in Table 2.

Table 2. Theme roles, markedness, patterns and examples

<table>
<thead>
<tr>
<th>Theme role and markedness pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Unmarked Subject (S(NP(<em>))(VP(</em>))(*))</td>
<td>The duke has given my aunt that teapot.</td>
</tr>
<tr>
<td>2 Existential there (S(NP(EX))(VP(<em>)(</em>)(VP(*)))</td>
<td>There were three jovial Welshmen.</td>
</tr>
<tr>
<td>3 Unmarked nominalization (S(SBAR(<em>))(VP(</em>))(*))</td>
<td>What the duke gave to my aunt was that teapot.</td>
</tr>
<tr>
<td>4 Marked Adjunct: Adverbial phrase (S(ADVP(<em>))(NP(</em>)(VP(*)))</td>
<td>Merrily we roll along.</td>
</tr>
<tr>
<td>5 Marked Adjunct: PP (S(PP(<em>)(VP(</em>)))</td>
<td>From house to house, I went my way.</td>
</tr>
<tr>
<td>6 Marked: Object as theme (S(NP(<em>)(VP(</em>)))</td>
<td>That teapot the duke has given my aunt.</td>
</tr>
<tr>
<td>7 Marked nominalization (S(SBAR*)(NP(<em>)(VP(</em>)))</td>
<td>What they could not eat that night, the Queen next morning fried.</td>
</tr>
<tr>
<td>8 Exclamative (S(ADJP(<em>)(VP(</em>)))</td>
<td>How dreadful she sounds.</td>
</tr>
<tr>
<td>9 Unmarked: WH-question (SBARQ(*))</td>
<td>Who wants a glass of wine?</td>
</tr>
<tr>
<td>10 Unmarked interrogative: yes/no (SQ(*))</td>
<td>Did you sleep okay?</td>
</tr>
<tr>
<td>11 Marked: Inverted clause (SINV(*))</td>
<td>On the right is it?</td>
</tr>
<tr>
<td>12 Unmarked imperative (S(VP(<em>)(VP(</em>)))</td>
<td>Turn it down.</td>
</tr>
<tr>
<td>13 Unmarked thematic equative (S(NP(NP*)(SBAR(<em>)))(VP(</em>)*))</td>
<td>The one who gave my aunt that teapot was the duke.</td>
</tr>
<tr>
<td>14 Marked thematic equative (S(NP(DT))(VP(SBAR(<em>))(</em>))</td>
<td>That is the one I like.</td>
</tr>
</tbody>
</table>

For instance, when given Example 3 in Table 2, (S(SBAR(*))(VP(*))(*)), as is the case with What the duke gave to my aunt was that teapot, the analyzer knows that the mood type is declarative because the clause begins with the syntactic tag.
S, the main clause, instead of SQ, an interrogative clause, or S (VP), an imperative clause. The SBAR clause is an embedded clause followed by a VP sister, and is therefore analyzed as Subject of the clause. Because the embedded clause is the first topical element as well as the first head phrase in the clause, it is the theme of the clause. Thus, the syntactic node of the theme is SBAR, the theme role of the theme is Subject, and the theme is unmarked. A marked counterpart of Example 3 is Example 7, (S(SBAR(*))(NP(*))(VP(*))), for *What they could not eat that night the Queen next morning fried*. The only difference is the intervening NP between SBAR and VP. Here, the SBAR is the Object and the theme whereas the NP is the Subject. As an Object is not the typical theme for the declarative mood, the analyzer determines that the theme is marked.

Once the analyzer has extracted all necessary information about the theme, it creates a report for human users. Figure 6 shows an example of the analysis report produced by the analyzer.

<table>
<thead>
<tr>
<th>Clause</th>
<th>Other problems may arise in this process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme</td>
<td>Other problems</td>
</tr>
<tr>
<td>Syntactic node</td>
<td>NP</td>
</tr>
<tr>
<td>Theme function</td>
<td>Topical</td>
</tr>
<tr>
<td>Markedness</td>
<td>Unmarked</td>
</tr>
<tr>
<td>Mood type</td>
<td>Declarative</td>
</tr>
<tr>
<td>Theme role</td>
<td>Subject</td>
</tr>
</tbody>
</table>

*Figure 6. Reporting theme information*

The report in Figure 6 reads as follows: the role of the topical theme here is the Subject of this declarative clause, which is an unmarked choice. Note that the writer could have made a choice that was more marked, e.g. by beginning the clause with the prepositional phrase, *in the process*. This choice would have reflected the writer’s intention to highlight the context in which problems may occur. In that case, the analyzer would have identified the theme as marked, as the use of the Adjunct as theme is an atypical choice for the declarative mood.

### 3.4 Technical details

Implemented in the Python programming language, the analyzer is freely available to the research community online. The analyzer requires Python 2.5 or later installed on a UNIX-like platform (including LINUX, Mac OS X, and UNIX), 1 GHz

---

5. The analyzer is downloadable from http://themeanalyzer.appspot.com. The script is licensed under the GNU general public license.
or higher CPU power, and a reasonable amount of memory (2GB or more) for optimal performance. The Stanford parser (Klein & Manning 2003) used for preprocessing is also freely available online. With the parse trees generated by the Stanford parser as input, the analyzer processes 5,000 sentences from the WSJ corpus in approximately 21 seconds at a rate of approximately 238 sentences per second.

4. System evaluation

The Theme Analyzer was developed using 264 theme-tagged model sentences found in published works in the field of Systemic Functional Linguistics (e.g. Eggins 2004, Halliday 1994, Thompson 2004). Specifically, the model sentences were used to develop and revise rules. For this reason, the model sentences are referred to as ‘training data’.

The test data used to evaluate the performance of the analyzer consisted of 500 sentences randomly selected from two large corpora of differing nature: 250 from the WSJ section of the Penn Treebank and 250 from the BAWE corpus. The WSJ corpus is a three-year collection of work from The Wall Street Journal with approximately 30 million words, and the BAWE corpus consists of 2,761 pieces of writing by British students, with each essay ranging in length from 500 to 5,000 words.

The sentences in the test data were manually annotated for thematic structure by two annotators. Inter-annotator agreement was assessed on a sample of 100 sentences (a subset of the test data), with 50 from the WSJ corpus and the 50 from the BAWE corpus. The themes of these sentences were identified independently by each of the two human annotators. Inter-annotator agreement on theme boundary identification was computed using the metrics of precision, recall, and F-score (Lu 2010), as given in the three formulae below, in which S1 and S2 denote the annotations of each of the two annotators. These metrics are widely used in the natural language processing literature to evaluate systems that perform identification tasks, such as word segmentation, which involves identifying word boundaries, and syntactic parsing, which involves identifying phrasal and clausal boundaries. In the precision and recall formulae, two themes in S1 and S2 are identical if they have the same boundary. The F-score, which is based on precision and recall, is


7. Kappa is commonly used to assess inter-annotator agreement on classification tasks in which annotators classify linguistic items into a certain number of mutually exclusive categories. However, the F-score is more commonly used to assess inter-annotator agreement on tasks involving the identification of boundaries of linguistic units.
the rate of inter-annotator agreement. The most informative measure in this case is the F-score.

i. Precision = Number of identical themes in S1 and S2 / Number of themes in S1
ii. Recall = Number of identical themes in S1 and S2 / Number of themes in S2
iii. F-score = 2 * (Precision * Recall) / (Precision + Recall)

In the process of manual tagging, one annotator identified 98 themes, whereas the other annotator identified 93 themes. Among these, 88 themes were identical, i.e. they had matching boundaries. The precision score was 0.90 (88/98), and the recall score was 0.95 (88/93). Based on precision and recall, the inter-annotator agreement was 0.92, which suggests that the two annotators had a very high rate of agreement on theme identification. Differences in the annotations were resolved via a discussion between the two annotators whereby they ultimately agreed on 99 themes.

The annotators also manually coded the 99 themes identified with the following information: theme type (e.g. topical or textual + topical), markedness (marked or unmarked), mood choice (e.g. declarative or interrogative) and theme role (e.g. Subject, Complement, or Adjunct). Across the 99 themes, the annotators agreed on 96 cases for theme type (Kappa = .66), 98 cases for markedness (Kappa = .95), 99 cases for mood choice (Kappa = 1.00), and 98 cases for theme role (Kappa = .96).

Following the assessment of the inter-annotator agreement, the two annotators each annotated another 200 sentences in the test set and then checked each other’s annotations. Discrepancies were again resolved through discussion. Altogether, the 500 manually annotated sentences served as the gold standard against which the performance of the Theme Analyzer was compared. In what follows, we report the results of the evaluation and provide an analysis of the errors committed by the analyzer.

4.1 Results of theme identification

The Theme Analyzer’s annotation of the 500 sentences in the test data was compared against the human annotation. The annotators identified a total of 488 themes in the 500 sentences, whereas the analyzer identified 468 themes. Among these identified themes, 438 were found to be identical, i.e. the annotators and the system agreed on the boundary of the theme. Agreement between the analyzer and the human annotators was computed using the same metrics of precision, recall, and F-score described above. The scores were 0.91 for precision and 0.95 for recall.

8. Similar to the interpretation of Kappa, the interpretation of the F-score is based on magnitude rather than statistical significance.
The F-score, based on precision and recall, was fairly high at 0.93, suggesting that the analyzer achieves a high degree of reliability in theme identification.

Among the 438 themes successfully identified by the Theme Analyzer, the Theme Analyzer agreed with the human annotators on 433 cases for theme type (Kappa = .97), 427 cases for markedness (Kappa = .96), 431 cases for mood choice (Kappa = .88), and 427 cases for theme role (Kappa = .97). These agreement rates indicate that the analyzer is very reliable in terms of identifying theme information.

4.2 Error analysis

One challenge for the analyzer identified through our error analysis was the emergence of new structures. Several structural variations that were not represented in the model sentences surfaced in the test data. Some examples of these new structures include reported speech, as illustrated in Example (3), and a T-unit with a lengthy subordinate sentence beginning with *as*, as illustrated in Example (4). In these examples, the themes are italicized:

(3) The reasonable first-year rates can be followed by increases of 60% or more if a covered employee files a major claim, *they* complain.

(4) *The report, released yesterday,* comes as Congress is considering a number of special tax breaks only three years after the sweeping tax-revision legislation abolished or curtailed many loopholes.

These structures were new (to the analyzer) and complex and were not successfully handled by the rules we designed. In addition to these challenges, there were some parser-level errors, including segmentation errors and parsing errors. These errors are attributable to the performance of the parser, not the Theme Analyzer per se. Segmentation errors occur when the parser incorrectly identifies the sentence boundary. In Example (5), the parser mistakenly considers the two sentences to be one sentence. Parsing errors, on the other hand, refer to cases where the parser fails to recognize the syntactic structure of a sentence. As Example (6) shows, instead of correctly recognizing *suffice it to say* as an adverbial clause, the parser erroneously identifies the clause as an independent imperative sentence.

(5) In addition, they will receive stock in the reorganized company, which will be named Ranger Industries Inc. After these payments, about $225,000 will be available for the 20 million common shares outstanding.
5. An example application

In this section, we illustrate how the automatic Thematic Analyzer is used to examine the differences in thematic selection between novice and expert academic writing. Our dataset consists of 2,761 student essays from the BAWE corpus selected to represent novice academic writing and 137 articles retrieved from Language Learning & Technology (LLT), a peer-reviewed scholarly journal, selected to represent expert academic writing. For this comparative analysis, we randomly selected 5,000 sentences from each of the two subsets of data, and analyzed the thematic structures of these 10,000 sentences using the analyzer. In what follows, we focus our discussion on four aspects among the thematic features that the analyzer reports: theme, syntactic node, markedness, and theme role. Table 3 summarizes the data and the results of the automatic theme analysis.

Table 3. Themes in student and expert writings

<table>
<thead>
<tr>
<th></th>
<th>Student writings (BAWE)</th>
<th>Expert writings (LLT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Mean length of sentences</td>
<td>18.59 words</td>
<td>26.63 words</td>
</tr>
<tr>
<td>Mean length of theme</td>
<td>4.09 words</td>
<td>6.14 words</td>
</tr>
<tr>
<td>Types of theme roles</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Types of syntactic nodes</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>
In terms of theme length, the results show that expert writers tend to produce longer themes than do student writers, as illustrated by the examples in Table 4. A closer examination of the data suggests that the elongated themes in expert writings can be partially attributed to expert writers’ systematic and strategic use of nominalization, which is a key feature of advanced academic writing (Schleppegrell 2001). Through nominalization, expert writers pack multiple pieces of information into a single theme. Moreover, nominalization also allows expert writers to condense the discussion from prior sentences into the thematic portion of a sentence and then develop the argument further in the rheme (Halliday 1994).

The results also reveal that compared with student writers, expert writers produce a greater degree of variation in the types of syntactic nodes. This suggests that expert writers have a wider repertoire of thematic choices, as exemplified in Table 5. As shown in Table 5, whereas student writers primarily produce NP themes, expert writers are much less dependent on NPs but engage substantially more diverse syntactic nodes in theme position, such as SBARs (subordinate clauses) and PPs (prepositional phrases). Within the same type of syntactic node, expert writers also tend to engage more complex structures than student writers. For example, in the case of NPs, expert writers are more likely to use nominalizations than students (as discussed above), whereas student writers are more likely to use single nouns. This tendency also extends to the case of PPs, as illustrated in Examples (7) and (8). In Example (7), a sentence from the BAWE corpus, the PP theme contains a single noun. In contrast, the expert in Example (8) comprises a PP theme that contains a rather complex nominal phrase.

(7) In my opinion, there are also many possible benefits of this research.

(8) In accordance with these major paradigm shifts in both general and applied linguistics toward acknowledging intonation as an indispensable component of language and communication, as well as with the advances in acoustic phonetic technology, this paper has three main goals.
Finally, the greater variation in the expert writers’ node choices, as compared to the limited choices made by students, seems to influence markedness and theme roles. The number of marked themes in the expert writings is 1,033, which is noticeably larger than 400 in the student writings. The frequent use of marked themes by experts can be explained in terms of, and related to, theme role.

### Table 5. Syntactic nodes of themes in student and expert writings

<table>
<thead>
<tr>
<th>Syntactic node of theme</th>
<th>Students (BAWE)</th>
<th>Experts (LLT)</th>
<th>Log Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun Phrase</td>
<td>4,379</td>
<td>3,472</td>
<td>105.02**</td>
</tr>
<tr>
<td><em>These perspectives are directed at […]</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepositional Phrase</td>
<td>339</td>
<td>924</td>
<td>13.58*</td>
</tr>
<tr>
<td><em>In authentic settings, culture and language are […]</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subordinate clause (a conjunction)</td>
<td>80</td>
<td>294</td>
<td>130.20**</td>
</tr>
<tr>
<td><em>As described in the last section</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent clause, nonfinite (participle and to-infinitive)</td>
<td>59</td>
<td>106</td>
<td>13.32*</td>
</tr>
<tr>
<td><em>Using item response theory, testers can […]</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verb Phrase</td>
<td>72</td>
<td>83</td>
<td>0.78</td>
</tr>
<tr>
<td><em>Find restaurants to please everyone</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others***</td>
<td>71</td>
<td>121</td>
<td>281.59**</td>
</tr>
</tbody>
</table>

Note. * denotes p < .001; ** denotes p < .0001; *** “Others” refers to low-frequency items, i.e. those that appear fewer than 100 times out of 100,000 identified themes. They are themes in clause complexes, adverbial groups as themes, predicated/comment clauses, mood adjuncts as interpersonal themes, themes in exclamative clauses, and themes in “yes/no” questions.

Finally, the greater variation in the expert writers’ node choices, as compared to the limited choices made by students, seems to influence markedness and theme roles. The number of marked themes in the expert writings is 1,033, which is noticeably larger than 400 in the student writings. The frequent use of marked themes by experts can be explained in terms of, and related to, theme role.

### Table 6. Theme role in student and expert writings

<table>
<thead>
<tr>
<th>Theme role</th>
<th>Students (BAWE)</th>
<th>Experts (LLT)</th>
<th>Log likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject as theme</td>
<td>4,198</td>
<td>3,390</td>
<td>86.20**</td>
</tr>
<tr>
<td>Adjunct as theme</td>
<td>339</td>
<td>924</td>
<td>281.59**</td>
</tr>
<tr>
<td>Adjunct (dependent clause) as theme</td>
<td>80</td>
<td>294</td>
<td>130.20**</td>
</tr>
<tr>
<td>Existential <em>there</em> in theme</td>
<td>139</td>
<td>47</td>
<td>47.57**</td>
</tr>
<tr>
<td>To-infinitive clause/gerund as theme</td>
<td>59</td>
<td>106</td>
<td>13.58*</td>
</tr>
<tr>
<td>Predicator</td>
<td>71</td>
<td>83</td>
<td>0.94</td>
</tr>
<tr>
<td>Complement as theme</td>
<td>46</td>
<td>94</td>
<td>16.80**</td>
</tr>
<tr>
<td>Others***</td>
<td>68</td>
<td>62</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Note. * denotes p < .001; ** denotes p < .0001.
As shown in Table 6, student writers (BAWE) have a very strong tendency to use a Subject as theme, whereas expert writers (LLT) make more balanced and varied choices of thematic roles. Also, the considerably more frequent use of Adjuncts, to-infinitives, gerunds, and complements as themes on the part of the expert writers compared with the limited use of these by the student writers also gives rise to a higher number of marked themes in the expert writings.

In sum, the comparison between the thematic choices made by student and expert writers brings to the fore the following three notable characteristics of student writings: the overuse of the Subject NP as the theme, a preference for unmarked themes, and less variation in the choice of thematic roles than in expert writings. By using the automatic Theme Analyzer, we can find out more about these characteristics through rapid data-based analysis.

6. Discussion and conclusion

This paper described a computerized system, Theme Analyzer, that automatically analyzes the thematic structure of English sentences. The system was developed using model sentences collected from the SFL literature, and it was tested using sentences randomly selected from the WSJ section of the Penn Treebank and the BAWE corpus. The test results show that the performance of the analyzer is reliable, as attested to by its high rate of agreement, 0.93, with human annotation. An analysis of the analyzer’s errors, that is, where the analyzer did not agree with human judgment, revealed two major problem sources, namely, thematic structures that lack clear definitions and parsing and/or sentence segmentation errors returned by the third-party parser. In an example application, the analyzer was used to analyze 10,000 sentences selected from the BAWE corpus and a collection of articles published in a peer-reviewed scholarly journal in order to identify differences in thematic structure between novice and expert academic writing. This comparative analysis revealed noticeable differences in student and expert writers’ thematic choices.

Our experience developing and applying the analyzer suggests that it has clear potential benefits for systemic functional linguists and corpus linguists as well as language educators interested in the functional approach to language and the use of corpora in language pedagogy. For systemic functional linguists, the program provides an efficient and reliable tool for performing automatic thematic structure analysis on large amounts of text data, alleviating them of the intensive labor required in manual analysis. For corpus linguists, the Theme Analyzer means they can automatically add a functional layer to corpus annotation and thereby draw on large-scale corpus data to examine information structure either on its own or in relation to other textual characteristics. For language educators, thematically tagged
corpora constitute a useful resource for showing students how meaning is created functionally as well as structurally to serve our communicative needs in real language use, which is an important goal of data-driven, corpus-based pedagogy.

Future research will focus on improving the performance of the analyzer and applying the analyzer to extended sets of data. First, the analyzer can be enhanced to handle the types of thematic structures that it currently does not understand, as discussed in the error analysis. This can be achieved by testing the analyzer against larger datasets and performing additional error analysis. Second, the analyzer can be applied on a large scale to texts written in diverse genres and by varying groups of writers. Analyses of this nature would allow us to answer such important questions as whether thematic choice differs systematically among writers with differing L1 backgrounds, whether it changes as English learners’ proficiency in the L2 improves, and whether it varies across variables such as genres and disciplines in academic settings, among others.

References


Authors’ addresses

Kwanghyun Park
Department of English Language & Literature
Myongji University
34, Geobukgol-ro, Seodaemun-gu
Seoul 120–728
Korea
kpark@mju.ac.kr

Xiaofei Lu
Department of Applied Linguistics
The Pennsylvania State University
301 Sparks Building
University Park, PA 16802
USA
xxl13@psu.edu