UNIVERSITY OF CALIFORNIA SANTA CRUZ

COMPUTATIONAL STYLE PROCESSING

A dissertation submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Foaad Khosmood

December 2011

The Dissertation of Foaad Khosmood

is approved:

_____

Professor Robert A. Levinson, Chair

_____

Professor Magy Seif El-Nasr

_____

Professor Marilyn Walker

_____

Tyrus Miller,

Vice Provost and Dean of Graduate Studies

CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF EXPERIMENTS AND DETAILED EXAMPLES

ABSTRACT

COMPUTATIONAL STYLE PROCESSING: ANALYSIS, CLASSIFICATION AND TRANSFORMATION OF NATURAL LANGUAGE STYLES

BY

FOAAD KHOSMOOD

Our main thesis is that computational processing of natural language styles can be accomplished using corpus analysis methods and language transformation rules. We demonstrate this first by statistically modeling natural language styles, and second by developing tools that carry out style processing, and finally by running experiments using the tools and evaluating the results. Specifically, we present a model for style in natural languages, and demonstrate style processing in three ways: Our system analyzes styles in quantifiable terms according to our model (analysis), associates documents based on stylistic similarity to known corpora (classification) and manipulates texts to match a desired target style (transformation).

In our model, we view style markers as the building blocks of styles. Style markers are features that can be extracted from the text which directly or indirectly reflect choices made by the authors. In order to perform stylistic analysis, we have developed a full marker taxonomy, a large library of marker instances and associated marker extraction routines.

To perform style-based classification, we extract a vector of markers from documents and corpora. Using nearest neighbor methods, we compare vectors to each other and calculate distances, and use the distances to perform supervised classifications. We derive a weighted distribution of markers for like-labeled document collections which we consider the mathematical representation of their collective style. We further experiment with varying the number of target classes and find a general decline in training convergence ability with increased number of classes. In addition we find that some markers are more persistently useful than others in our experiments.

We also demonstrate machine transformation of texts with detectable styles. Much like Statistical Machine Translation (SMT) from one language to another, we demonstrate that machine transformation of one style to another is achievable. A number of monolingual text-to-text transformation routines (transforms) must be available and applied intelligently and systematically to a document. We have developed a small number of such transforms to demonstrate the possibilities for style transformation. We show that a finite number of transformations on a document will lead to a change in the document's overall style.

Our experimental results support our thesis that automatic style processing is empirically possible. As a final step we demonstrate statistically significant correlations between our model of style and the common human understanding of style with a 19 subject user study.

to Dr. Kholood Hassan and Luna K. Hassan

ACKNOWLEDGEMENTS

I thank Professor Robert Levinson for his open mind, refreshing ideas, emphasis on scientific rigor and for his dedicated mentorship of me on this dissertation and far beyond.

I thank Professors Marilyn Walker and Magy Seif El-Nasr for their valuable time, advice and for pushing me to explore directions I was clearly not comfortable with at first. I thank Professors Michael Mateas, Kip Tellez and Noah Wardrip-Fruin for taking a chance on this idea and providing valuable feedback during my advancement. UCSC Professors Ira Pohl, Jim Whitehead, Arnav Jhala, David Helmbold and Patrick Tantalo have earned my gratitude for their great support and advice throughout the years. Finally, I have to thank my wife and partner, Dr. Kholood Hassan for her incredible patience, support and sacrifice. I could not have accomplished any of this without her active participation.

Portions of this text have been drawn from previous publications [Kho06], [Kho08], [Kho09], [Kho10], [Kho10-2] and [Kho11].

1.0 INTRODUCTION AND MOTIVATION

Style is an integral part of natural language in written, spoken, or machine generated forms. Natural language styles are understood, mimicked and transformed by human agents with ease. We believe that like natural language processing (NLP) in general, natural language styles can also be processed, i.e. recognized, generated, classified and transformed computationally. To test this belief, **we build a modular, extensible prototype system that automatically performs style-based analysis, classification and transformation on written language**. We call the system modular because we would like to accommodate an open set of stylistic markers, language operators and evaluation and reasoning methods, able to work together in a large variety of combinations in order to deliver the best possible results for style processing tasks.

1.1 SCOPE OF THE WORK AND CONTRIBUTIONS

The present study builds on research from several distinct communities: The first is Computational Stylistics, specifically digital authorship studies and authorship attribution; Second, Statistical Machine Translation (SMT) including monolingual text-to-text translation, and paraphrasing; And Thirdly, Artificial Intelligence, Machine Learning (ML) as applied to Natural Language Processing (NLP). Among these, the first community (Computational Stylistics) would be the closest natural base for our

contributions in style processing, even though traditionally few text manipulation and generation works have been done in this area.

This chapter serves as a quick introduction to the major concepts of the present work (1.2), an overview of our style processing prototype system (1.3), the use cases of the system (1.4), future applications of the technology (1.5) and two small, manually calculated examples to further demonstrate the approach (1.6). Next we list the overall hypothesis of this work as well as five derivative hypotheses along with tests for each (1.7). This chapter is meant to familiarize the reader with the concepts and provide a non-technical overview of the work. Lastly, the organization of this document and a preview of later chapters are discussed (1.8).

1.2 MODELING STYLE

Broadly speaking, style is a *way* of doing something. By implication that same thing can be done in more ways than one. Style is often associated with a person, role or entity. For example, in the game of chess, an *opening* refers to the first few moves of a player. If we define *chess style* as having a distinct opening, we could identify the opening of a particular player's chess style just by examining the record of a series of games by the player whom we already associate with having that style. The opening would be one recognizable factor in the style of chess play being observed.

All natural languages and even most artificial languages can be associated with one or more "styles." There is no universally acknowledged definition of what specific elements constitute "style" when it comes to language. The notion is inherently

imprecise, rendering any decent contextual definition necessarily broad. Linguists often use terms like “dialect” and “register” as concepts that fit inside the larger notion of style. Dialects are socio-graphical and strongly associated with actual agents. Register is a linguistics concept describing language along the axes of field (subject matter), tenor (formality and social relationships) and mode (medium of communication). Dialect and register are examples of concepts that fit within the larger category of style.

On the other hand, completely non-linguistic and typographical choices of written language are often called style as well. Examples include font size, color, text decorations, visual emphasis, indentation and usage of non-linguistic symbols.

In the background section below, we discuss how we adapt and appropriate a suitable definition of style for this work.

1.2.1 DETECTING STYLE

Given that styles are omnipresent in language and they can be easily detected by humans [Wal08], we should be able to at least partially define a series of features detectable in each distinct style [Whi04]. These are some of the same features that convince a human examiner to assign a distinct style label to a piece of text, considered either on its own or in comparison to one or more other texts. Fortunately, there is a large body of research in AI, ML, NLP and Computational Stylistics that have considered this type of algorithmic feature selection and text classification in depth [Har68][Fak01][Kes03][Kar04][Juo06]. In addition, work in linguistics, sociolinguistics, corpus linguistics, literary studies and language studies provide a rich repository of

analysis about language variation that can be used to augment aforementioned algorithmic and mathematical methods [Buf21][Cry69][Car88][Ris94][Van03][Sid10]. We hope to employ this literature to derive ever richer and more complex style markers that could deliver more precise classifications.

1.2.2 STYLE-BASED TRANSFORMATION

Style-based transformation is the idea of linguistically altering a piece of text so it exhibits the characteristics of other pieces of text, which are associated with a style. It is essentially a text-to-text language translation operation similar to SMT except that instead of translating from one language to another, it would translate from one style to another. As with SMT, the meaning of the original message has to be preserved as much as possible.

While text-to-text computational paraphrasing work exists, it is almost always for a specific user-centric purpose such as summarization or simplification [Sid10], not style manipulation. Overall, the most likely areas one can expect to encounter this particular exercise are literary analysis [Hoo99], education [Wal80], writer development [Hei00] and even in those areas it is a task for human agents, not artificial ones.

However some loosely related concepts and technologies exist in AI and NLP, which will be helpful to us. Natural Language generation can perhaps be considered "half" the problem [Lan94][Ger00][Lan02][Mai10]. Once the intention or the intended meaning of the utterance is known, NLG techniques can produce multiple stylistic

variants of the text. Some productivity tools and computational writing analyzers are often implemented as rule-driven expert systems capable of suggesting paraphrases in order to improve the quality of writing or alter it in various ways. Although the goal of these paraphrase routines would not be the same as what we have in mind, they nevertheless do represent automatic language alteration tools.

1.2.3 CLASSIFICATION-TRANSFORMATION LOOP

We contend that transformation is intimately related to classification and vice versa. Successful transformation depends on precise classification in the sense that a given classifier ultimately determines whether or not the transformed text exhibits enough markers to be considered as one of the texts of the target style. Style markers that help the classification of text, in turn, provide the basis for style transforms. This conception allows for an elegant "symbiotic" growth in sophistication for both sub-systems. For the design of our prototype, we use this interdependence to focus our efforts where they would produce the best results. The interdependence between classification and transformation can be illustrated by the following example:

We imagine a style we could call "Shakespearian," derived from a corpus consisting of several of Shakespeare's tragedies. We find that our classifier can easily distinguish between Shakespearian corpus and a modern English corpus just by examining the presence of early modern English pronoun forms (such as "thou", "ye", or "thine"). This gives us a clear path to build a transform that simply converts all the modern pronouns in a given text to the Shakespearian English equivalent. The

transformed text would now be classified in our own system as a “Shakespearian.” However, when read by human experts, the piece would be far from Shakespearian; after all, there is more to Shakespeare than pronouns. We conclude that the classification is in fact incorrect because the style of Shakespearian Tragedy was grossly underspecified. The corrective action, then, would be to add more style markers that specifically discriminate between the once-transformed text and the Shakespearian corpus. We can do this by observing more unique features of Shakespeare. Examples are: iambic pentameter, presence of characters, playwright-style divisions (i.e. acts and scenes), distinct vocabulary, motifs of violence, death or betrayal, etc. Each of these observations provides two things for us simultaneously:

1. The basis for a marker that can be analyzed by our classifier
2. The basis for a transform that is capable of altering text such that the result exhibits the phenomenon in question

The second task requires progressively sophisticated linguistic text-to-text operators. To be sure, not all style markers could yield to equivalent transformers without also producing significant side effects. Changing general text to iambic pentameter, for example, would be extremely difficult if not impossible. But the point is that each marker embeds within it a related transform. And each transformation process produces text that might still be lacking the desired style as observed by humans, prompting for more complete style specification through finding better markers. This is a spiral that can continue as long as there are creative ways to make markers and transforms or until such time as a reasonable human observer could

confidently decide that the latest transformed piece of text is "Shakespearian enough." In essence, we can systematically leverage a human's vastly superior familiarity with numerous style categories to inform a machine's classifier that works based on limited number of known styles.

1.3 EXTENSIBLE, MODULAR, HETEROGENEOUS SYSTEM

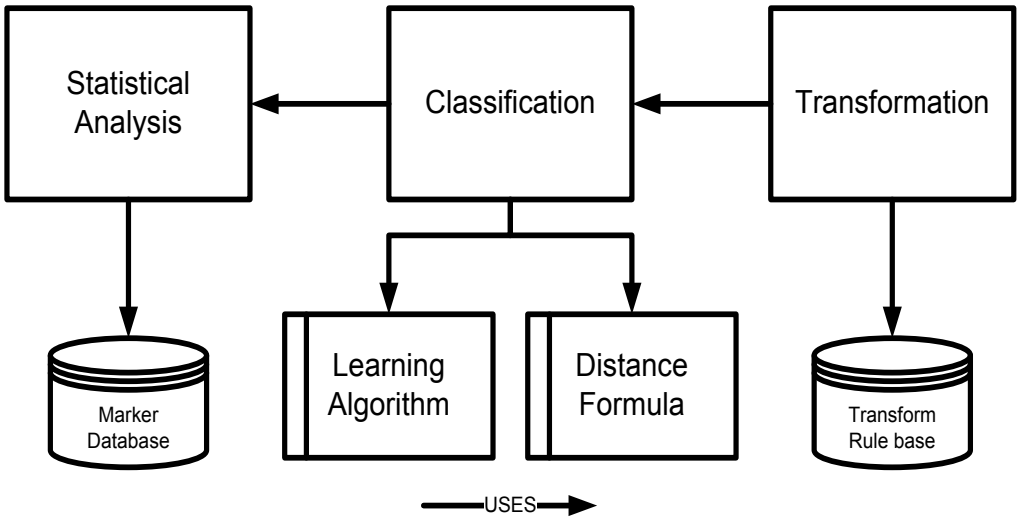

Figure 1. Major components of a style processing system

We are interested in building a system capable of performing stylistic analysis, classification and classification-aided transformation as described above. But we also want the system to have maximum flexibility to consider a large set of markers (or intelligently chosen subset) and employ a large number of transforms. We would like the system to take advantage of markers and transforms that have been developed from a wide variety of sources for different purposes and apply them to tasks as required. Lastly, we would like the system to be easily augmentable.

The above requirements necessitate the use of well-defined layers of abstraction. A standard interface for style markers and style transforms is necessary in order to exploit heterogeneous operations. In addition to markers and transforms, the classification, transformation and evaluation methods themselves should be extensible. The system should allow plugging in of off-the-shelf machine learning and clustering algorithms for its classification section. Similarly, optimizers and planning agents could be enhanced to make transformation more efficient.

1.4 USE CASES

We outline some major use cases for a system described above. There are fundamentally three basic operations as shown in Fig. 1: Analysis, Classification and Transformation. Each of these can be mixed and matched to achieve more complex goals. Fig. 2 below describes some of the style processing use cases we support with this system.

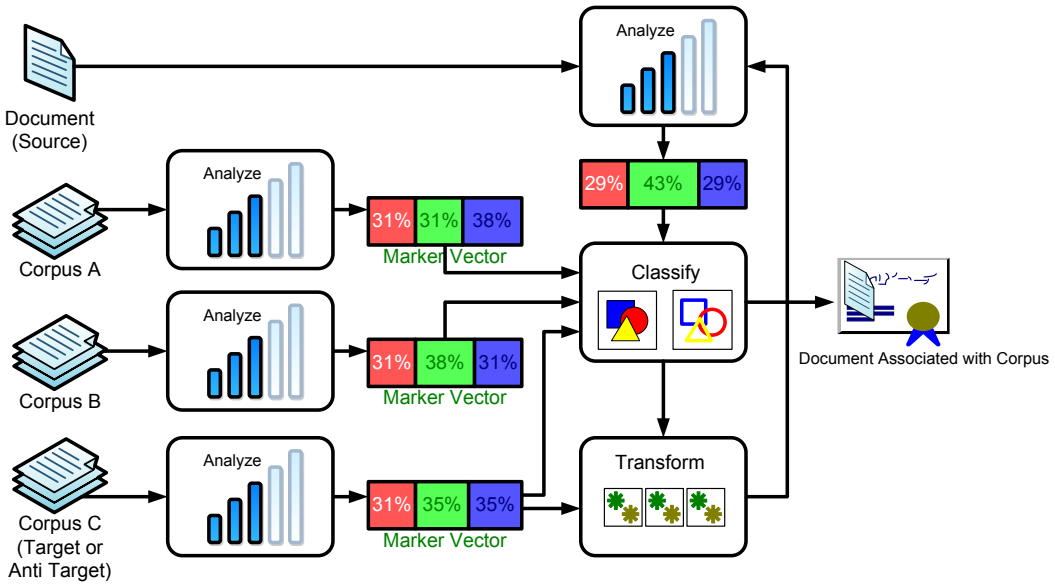

Figure 2. Supported style processing use cases

1.4.1 Analyze a document or corpus

Example: "What is the average sentence length in the U.S. Constitution?"

A document such as "Source" or a corpus can be analyzed. The analysis consists of identifying all the enabled markers, running the marker extraction routines and storing the results of each marker *instance* in the marker array. Each marker also has production parameters and quantifiers settings that have to be checked against. The settings generally control how many floating-point values are extractable from the same concept-family-marker entry.

1.4.2 Generate a similarity matrix between multiple corpora

Example: “Which two of the Brown Corpus categories are furthest apart?”

Our assumption is that distinct corpora have a relatively unique detectable style. Between two or more corpora, however, there can be similarity values calculated. Our system can accept multiple corpora and generate a two dimensional matrix where the distance between every corpus and every other corpus is represented. Such a matrix can provide insights as to the differences in style. The closer the distance between distinct corpora the more difficult it is to for the system to distinguish between them and classify documents correct. However, transformation between two close styles should be easier to accomplish than two distant ones.

1.4.3 Associate a document with a corpus based on style

Example: “Is *King Lear* a comedy or a tragedy?”

In this case, the source document, and two or more corpora are each analyzed with the same-size marker array extracted from all. The classifier system will accept all the corpus marker arrays; normalize their content and train on them for classification. A weight vector is produced, reflecting the relative importance of each marker for classification purposes. After sufficiently close convergence is achieved, the new document is examined and its weighted marker array compared against that of the corpora. The document is associated with the corpus with the closest distance.

1.4.4 Stylistically alter a document

Example: “What would my abstract look like with many more passive sentences?”

The library of transforms is available to be used on a document to alter its stylistic signature. Transforms can be applied to a document using handcrafted parameters and the process can be repeated.

1.4.5 STYLE OBFUSCATION USING AN ANTI-TARGET

Example: “Can I make this email sound less like a typical one from me?”

This is the situation where a document which is initially part of a corpus (Corpus A) is transformed, analyzed and compared against the collection. If the original association to Corpus A has now been broken (i.e. the new transformed document is no longer classifiable as a member of Corpus A), then the document is said to have been stylistically shifted away from Corpus A. This shift can be specifically focused to be away from Corpus A that is called the “anti target.” A distance formula is used to compare the document against the anti target and select for transformations that maximize that distance.

1.4.6 STYLE TRANSFORMATION USING A TARGET

Example: “I would like to change my essay to read a bit more like my favorite New York Times columnist.”

In this scenario a document is to be transformed and stylistically associated to one of the corpora from the training set. This document does not necessarily need to be part of any existing corpus. Through a systematic application of transforms, the system moves the document gradually closer to a target corpus, until such time as the document can be objectively associated with the target corpus, or the system runs out of transformation options and the exercise fails.

1.5 APPLICATIONS

Some applications that we foresee being able to make use of our system include digital forensics and plagiarism, authorial style obfuscation (for privacy purposes), style based search, authoring tools for development of narratives and personalities in writing and digital entertainment, machine translation, writing educational tools and adaptable computer agents for human interaction.

1.5.1 DIGITAL FORENSICS AND PLAGIARISM DETECTION

The classifier part of the system could be used to identify and match the style of a text to the style derived from examining corpora belonging to two or more suspects. Similarly in plagiarism detection, the system could match the style of the author being plagiarized against the plagiarized text and even provide the style markers that can serve as evidence for plagiarism.

1.5.2 STYLE OBFUSCATION

For a variety of reasons people may wish to intentionally obfuscate the style of their own writing. Style remains an easy way to recognize an anonymous piece of writing. Our tool could facilitate automatic stylistic transformation of a text toward an obscure style, or "away" from the stylistic observations made off the authors typical writing.

1.5.3 STYLE BASED SEARCH

Style based search proposes a new kind of mass-document search (such as Internet) where instead of supplying words and phrases as search terms, one would upload an entire document and the engine returns other documents with similar style within certain tolerance levels.

1.5.4 EDUCATIONAL AND PRODUCTIVITY TOOLS

Productivity software titles like Microsoft Word increasingly have integrated grammar and style tools to aid in writing and suggest paraphrases. With a style transform tool such as the one we are proposing, the user could have more options and flexibility. Instead of evaluating along just one axis (the generic "good writing style" that MS Word tools are coded for), the user could experiment with multiple known styles, or derive a new style from a corpus to be examined by the tool.

1.5.5 AUTHORING TOOLS AND DIGITAL ENTERTAINMENT APPLICATIONS

In addition to the productivity software functionality, there could also be more sophisticated author-centric tools, taking advantage of stylistic transformations that could aid in character development, narration and expository passages. Authors could get ideas by automatically paraphrasing their writing along stylistic dimensions. Characters in a dialogue could acquire different personalities detectable by their speaking styles. We see application to digital entertainment whereby written passages or character utterances could be dynamically transformed to some desired target using

some of our methods. This application could possibly free up game writers and designers from having to rewrite multiple versions of the same text in different styles.

1.5.6 Better Machine Translation (MT)

One of the problems inherent in MT is stylistic difference in various languages. After (an often literal) translation, the resulting text may be intelligible but still not in the form that the speakers of the target language are used to. A transformation between "post-MT style" and "native style" could solve this problem.

1.5.7 Interactive agents

Robots and digital agents that interact with human users through text (or speech directly synthesized from text) have often hard-coded communication styles. To take a simple example, Microsoft Office's "Clippy" was an interactive help agent that appeared in strategic moments and attempted to solve a perceived problem through a kind of text-based dialogue with the user. We could imagine more sophisticated agents with unique personalities each communicating in different styles. Ideally, such agents could be trained to morph their communication style to one best suited for the user. The training could come from analyzing user's existing writing pieces or emails. A stylistic profile of the user could be derived that could form the basis for the agent's personality, or at the very least, automatically select the best choice among multiple static personalities supplied.

1.6 DETAILED EXAMPLES

In this section we discuss in depth a number of examples of various complexities denoting the kind of work a fully functional system could perform. It is important to note that all scenarios in this section are hypothetical as a fully functioning system does not yet exist to perform them. However, given some assumptions, they should be practical to support a derivative system build on our prototype.

1.6.1 MODERNIZING SHAKESPEARE EXAMPLE

In this example, we demonstrate how a stylistic processing system could work to convert a few lines of Shakespeare to a more modern style.

Source *S* is in the style of Shakespearian Early Modern English, as profiled by Act I of Shakespearian play Hamlet. Style *T* is Modern American English as profiled by processing corpora of literary works in the 20th century written in modern American English. Several full-length novels by Tim O'Brien, Don DE Lillo and Toni Morrison should be enough to profile the style.

Marker-detector M1 measures the density of modern English pronouns in a piece of text. Style *T*, consisting of modern works, has more occurrences of these pronouns. Thus the relation M1(*T*) > M1(*S*) holds at the outset.

Operator P1 modernizes archaic pronoun forms such as "yea", "thou" or "thine", through strict lexeme substitution.

Sentence A, a part of S, is this line from the Shakespearian play *Hamlet*:

This above all: to thine own self be true,
And it must follow, as the night the day,
Thou canst not then be false to any man.

Operator P1 is applied to sentence *A*. Under specific implementation choices, P1(*A*) results in this sentence:

This above all: to your own self be true,
And it must follow, as the night the day,
You canst not then be false to any man.

B is assigned to the outcome, i.e. *B*=P1(*A*). This operation yields the following: M1(*B*) > M1(*A*) which means the text is now closer to target style *T* along the dimension measured by M1 marker. A whole family of such language-specific operators can be employed in the same fashion, maximizing other metrics measuring modernity of language use.

After successive applications of many transformer functions, the result may be like this for *B*:

This is the most important of all: be honest with yourself,
and then it follows, just like the night follows the day,
that you cannot be dishonest to anyone.

In addition we wish to maximize a new metric, M2, which measures the average number of words per sentence in a piece of text. It holds that M2(s) > M2(t). Operator P2 is designed to split large sentences into two, whenever possible. Applying P2(B) would yield the following:

> *This is the most important of all. Be honest with yourself. And then it follows, just like the night follows the day, that you cannot be dishonest to anyone.*

Operator P2 split the sentence into 3 sentences by simply converting colons and conjunctions to sentence boundaries. The result is that M2(B) > M2(P2(B)), in other words, the transformed source has moved closer to the target style after applying two transformations.

1.6.2 A TRIVIAL 2-MARKER SCENARIO IN DEPTH

In this scenario, we manually follow the actual steps in the system for a trivial transformation. We are given a corpus, T, which can be described as modern, formal, factual writing with short sentences. We only have two markers in this scenario, specified with two corresponding marker-detector functions, and two transform functions. The marker-detector functions are as follows:

1. M1() returns inverse of the average words per sentence (i.e. sentence/word).

2. M2() returns a ratio of overall instances of single digit numbers written using Arabic numeral symbols (i.e. “1”, “2”, etc.), to the total number of single digit numbers written in the text.

We are also given 2 transform functions:

1. X1() uses some symbolic linguistic techniques to split large sentences into two or more smaller ones.
2. X2() replaces instances of a number written in English to its equivalent in Arabic numeral symbols.

Finally, we are given the following text from a recent new story as source text (*S*):

> *Canada's opposition parties have reached a tentative deal to form a coalition that would replace Prime Minister Stephen Harper's minority Conservative government less than two months after its reelection, a senior negotiator said on Monday.*

This text contains 35 words and one of them is a number (“two”). A Simplified transformation run through the system is the following:

1. Analyze the target corpus, T (for this example we specify target corpus characteristics)
    a. M1(T) = 0.2
    b. M2(T) = 0.8
2. Analyze the source text, *S*
    a. M1(*S*) = (1/35) = 0.0286
    b. M2(*S*) = (0/1) = 0.0

3. For comparison and evaluation, a simple root mean squared error (RMSE) is calculated between the *S* and *T* vectors.
    a. RMSE(*S,T*) = $\sqrt{((0.2 - 0.0286)^2 + (0.8 - 0)^2)} = 0.8182$
    b. The system determines the RMSE distance is too high to constitute a match
4. Operator X1 is applied to source text, *S*
    a. Operator X1(*S*) converts the text to version S_2 (inserted words and punctuation are underlined)

    Canada's opposition parties have reached a tentative deal. The deal is to form a coalition. The coalition would replace Prime Minister Stephen Harper's minority Conservative government. This is less than two months after its reelection. A senior negotiator said it on Monday.

5. S_2 is analyzed
    a. M1(S_2) = ((1/8)+(1/7)+(1/11)+(1/9)+(1/7))/5 = 0.1225
    b. M2(S_2) = (0/1) = 0.0
6. New RMSE is calculated
    a. RMSE(S_2,*T*) = $\sqrt{((0.2 - 0.1225)^2 + (0.8 - 0)^2)} = 0.8037$
    b. System determines RMSE is still too high so we continue.
7. Operator X2 is applies to source text, *S*
    a. Operator X2(S_2) converts the single instance of "two" to "2" and produces S_3.

    Canada's opposition parties have reached a tentative deal. The deal is to form a coalition. The coalition would replace Prime Minister Stephen Harper's minority Conservative government. This is less than 2 months after its reelection. A senior negotiator said it on Monday.

8. S_3 is analyzed
    a. M1(S_3) = ((1/8)+(1/7)+(1/11)+(1/9)+(1/7))/5 = 0.1225
    b. M2(S_3) = (1/1) = 1.0
9. New RMSE is calculated
    a. RMSE(S_3,T) = $\sqrt{((0.2 - 0.1225)^2 + (0.8 - 1.0)^2)}$ = 0.2144
    b. At this point the system can either determine that this RMSE number is within tolerance levels of the target corpus, or it can equally decide that since the transform functions have been exhausted S_3 represents a best effort transformation for this problem.

Two aspects of this scenario are worth noting. First, the two marker-detectors M1 and M2 were considered with equal weight distributions. This does not have to be the case, the analyzer routine can assign alternative weightings to these markers, emphasizing or deemphasizing one of them. A new attribute could accompany each marker as a measure of relative importance. But since we have yet to find a good theory of indicating how important each marker should be, we continue to keep them all equally weighted.

Second, operators X1 and X2 were applied successively to the source text. The final S, (S_3) was therefore given by S_3 = X2(X1(S)). However, there are four possible orderings of these two operators: X1(S), X2(S), X1(X2(S)) and X2(X1(S)). What we did in this example is a sequential application of operators without any feedback considerations. But we could easily change the planning algorithm to perform exhaustive search (try every combination) or a greedy algorithm (apply the next step

with the highest pay off), or a genetic algorithm with the distance as a fitness function or any other algorithm we wish. This sub-problem is NP-complete and not really the main concern of this work. However we do allow for user specification of this algorithm.

1.7 LIST OF SPECIFIC HYPOTHESES AND TESTS

This section lists the main hypothesis of the work and the five specific hypotheses, each corresponding to an important component of the main.

Main hypothesis: Computational processing of natural language styles can be accomplished using corpus analysis methods and language transformation rules.

1. Style Markers and Classification

1.2 Elaboration: We define style as linguistic choices available to express the same meaning [Wal80]. Many existing systems privilege one (or few) markers in their style analysis. Many existing approaches fall into what J. Rudman [Rud03] has called "the problem of cherry picking", in that they use situation-specific markers and corpora for their analysis and show no proof of universality of the approach. Combining markers is difficult because there is no established super-set or even an established categorization and parameterization of these markers.

1.3 Hypothesis 1: Multiple combined markers perform better than one (or few) in attribution problems.

1.3.1 Test: Compare performance of single-marker systems with hypothetical multiple-marker variants of the same system.

1.3.2 Test: Generate own marker organization allowing for combination. Use it to build multi-marker attribution/classification system. Show broad attribution performance gets better with more markers.

1.4 Hypothesis 2: Some markers are more "universal" than others. They stand out in wider array of problems and are consistently more influential than others.

1.4.1 Test: Show experimentally that as the number of corpora/styles grows, some markers are relatively persistent (i.e. consistently highly weighted) across attribution problems.

1.5 Hypothesis 3: Machine classification based on our statistical model of style correlates with human classification based on style (common understanding).

1.5.1 Test: Human evaluators to classify some documents, given the same corpora and classes, and show correlation with machine classifiers.

2. Style Transformation

2.2 Elaboration: Transforms are monolingual text-to-text paraphrasing rule-based systems strategically applied to alter stylistic signature of texts, often individual sentences. Transforms tend to be linguistically specific. A large collection of them is necessary to yield more comprehensive results. We show the stylistic impact of some novel transforms, created by ourselves, combined with ones created by others in the community.

2.3 Hypothesis 4: Style shifting at the small unit level can be used to achieve style change at the larger, document level.

2.3.1 Test: Create a system to strategically apply style transforms given a library of them. Use human evaluators to verify that system unit-level transforms are a) grammatically correct and b) semantics preserving. If (a) and (b) are met, then we have accomplished unit-level style shifting.

2.4 Hypothesis 5: Incremental transformation from one style to another can be accomplished with language-specific transforms. Some style-to-style transformations are easier achieved than others, depending on the library of valid markers and transforms available to the system. Transformation can be targeted (toward a specific style) or anti-targeted (away from a specific style).

2.4.1 Test: Create libraries containing a variety of transforms and markets from original contributions and several different sources in the literature. Use the same style classification method in a typical classification exercise, and show that it will re-classify a document after a series of transforms have been applied to it. Demonstrate both targeted and anti-targeted varieties. Successful machine transformations to be verified by humans, given the same corpora and classes, in a user study.

Thus, by proving the hypotheses above, we will have demonstrated to varying degrees, that fundamental style processing operations are achievable. These operations include style marker analysis, style recognition, style classification, style transformation and we will have discussed some feasible methods to accomplish them. We revisit these hypotheses and determine how well they were achieved in Chapter 7, *Evaluation*.

1.8 THE ORGANIZATION OF THIS DOCUMENT

In the Background and Related Works chapter (2), we provide a survey of how various related disciplines have understood and worked with language styles. A historical overview, as well as contemporary work in style analysis is provided. Although most of the background and previous research is cited in this chapter, other related works specific to markers, classification and transformation are usually cited within those respective chapters.

Chapter 3 is dedicated entirely to markers, their operation and our proposed taxonomic representation. Chapter 4 covers style-based classification with full details of our approach. Chapter 5 lists the specific style transform functions used in our work, along with theory, design experiments and performance analysis for each. Chapter 6 details our user study (complete with discussion of its elements and how they were developed) as well as the results.

In Chapter 7, "Evaluation", we revisit the foundational assumptions and hypothesis that we presented in section 1.7 and conduct a short and factual assessment of all the hypotheses tested.

2.0 Background and Related Work

Writing style is difficult to define and indeed has no precise enough definition for computational purposes. In this section, we begin by familiarizing the reader with types of text classifications available. We discuss historical approaches toward style, as encountered throughout the literature space in multiple disciplines. We further discuss style markers used in mainly classification or text analysis research in order to provide a base for using them in our system. Next we cover relevant work in style markers (relevant to Chapter 3), authorship attribution systems (relevant to Chapter 4), paraphrasing (relevant to Chapter 5) and style obfuscation (relevant to Chapters 5 and 6).

2.1 Text classifications: genre, text type and register

Classification of written works for referencing and archiving purposes was perhaps the first area of human activity where the question of style may have been approached.

In [Moe01] a good background is give on how text type, genre and style relate to each other in linguistics. The author begins by describing genre as non-linguistic classification of written work in the tradition of Aristotle. When the designers of the Helsinki Corpus [Ris94] wanted a definitive classification of all English text, they found traditional genre as was used in the field of literary studies inadequate. They felt they

had to add classes of non-literary texts such as "letter, proceeding, trial" [Moe01] in addition to a catchall "none of the above" category. Thus [Ris94] called their classification variable "text type."

Both text type and genre are non-linguistic classifications, but each of their classes exhibits high correlation to one or more linguistic features. As [Moe01] describes:

> *Examples of such correlations are that narrative text types contain past tense verb forms; biographies are written in the third person singular, etc. The* ***set of linguistic features which characterizes a particular text type or genre*** *is referred to as the text type style or genre style. Some linguists also use the expression 'register' with this meaning.*

This gives a definition of "register" as a purely linguistically defined category. In [Bib89], Biber was able to derive purely linguistic text types by aggregating and clustering linguistic features. Biber identified 67 features, based on previous studies that had associated them with one or more genres. Through calculating various combinations of co-occurrence frequencies of these features, he produced 8 "text type" categories derived using automatic clustering algorithms.

2.2 Defining style and stylistics

Andreas Jucker [Juc92] likened "style" to what de Saussure calls "parole" and what Chomsky calls "performance." Jucker calls style a "comparative concept" describing relative differences between "texts or discourses" and in some cases between

text/discourse and "some kind of explicit or implicit norm" [Juc92]. He also contrasts the popular view of the term "style" with the notion in the traditional stylistics.

> *There is a lay notion of the concept of style, which equates style with the elevated and aesthetically pleasing forms that are used, for instance, by celebrated authors in their writings. Some newspapers, accordingly, are claimed to "lack style" altogether.*
>
> *This is of course not what traditional stylistics takes style to be.* ***Every single text has got a style as far as it has formal properties that can be compared with those from other texts.*** *A stylistic analysis will try to single out those features that help to distinguish the texts under comparison. One particular feature may occur in only one text and not the other, or it may appear with a frequency that is appreciably different from one text to the other*. [Juc92, page 12]

2.2.1 STYLE AS HABITS, TENDENCY AND CHOICE

Stylistics has been called a conceptual successor to the ancient Greek concept of rhetoric [Bib89]. One of the ways in which Crystal and Davy define style is as "a selection of language habits the occasional linguistic idiosyncrasies which characterize an individual's uniqueness" [Cry69].

Walpole calls stylistics an "offshoot of linguistics" which "takes the position" that "style is a deviation from normal language usage" [Wal80]. [Kar04] contextualizes style by calling it a "consistent and distinguishable tendency" to make linguistic choices.

Fish, in his essay "What is stylistics and why are they saying such terrible things about it," called stylistics simply the "application of linguistics to the study of literature" [Fish81]. Simpson writes in *Stylistics: A resource Book for Students* that "Stylistics is a

method of textual interpretation in which primacy of place is assigned to language" [Sim04].

2.2.2 STYLOMETRY AND STYLISTIC MEANING

George Heidorn of Microsoft is probably closer to the lay definition of "style", when he defines "style checking" as checking for deviations from "good writing style."

Stylometry is a term that is sometimes treated as a synonym for *stylistics* but occasionally treated as having a narrower scope than the latter, typically applicable to just one author or corpus. McEnery and Oakes describe stylometry as "an attempt to capture the essence of the style of a particular author…" [McE00].

Whitelaw and Argemon in [Whi04] propose a definition for "stylistic meaning."

> *We provisionally define* stylistic meaning *of a text to be those aspects of its meaning that are non-denotational, i.e., independent of the objects and events to which the text refers* [Whi04].

This definition raises one of the central questions concerning style: style's relationship with other concepts such as meaning or personality. While it may be more functional to think of style as inherently separate from content, personality and meaning, or only affecting the manner of delivery rather than the message itself, most experts agree that style is not disjoint from meaning entirely.

2.2.3 STYLE AND PERSONALITY

As far back as 1753, George-Louis Leclerc de Buffon declared that "Style is the man himself," [Buf21] implying an inseparability between style from personality of the author. While saying style is unique to an individual forms a solid theoretical basis for authorship attribution that we examine below, it does suggest manual style transformation or style imitation may not be possible. Similarly, John Middleton Murry wrote that "style is not an isolable quality of writing; it is writing itself" [Mur22]. Simpson admits that style affects meaning, but it is not meaning in its entirety:

> *While linguistic features do not of themselves constitute a text's 'meaning', an* ***account of linguistic features nonetheless serve to ground a stylistic interpretation*** *and to help explain why, for the analyst, certain types of meaning are possible.*
>
> *These 'extra-linguistic' parameters are inextricably tied up with the way a text 'means'. The more complete and context-sensitive the description of language, then the fuller the stylistic analysis that accrues.* [Sim04, page 2]

Walpole suggests that style and discourse may not be so easily separable when she writes that "We no longer feel comfortable viewing style as the dress of discourse, the external and changeable garb for ideas…" [Wal80]. However, as we write below, Walpole's model of "style as option" does allow for both extra-linguistic analysis and even the possibility of style transformation.

2.2.4 SUMMARY OF THE ENCOUNTERED NOTIONS OF STYLE

Table 1 summarizes the definitions and commentaries discussed in this section. Ours is by no means an exhaustive list of style definitions. Many thinkers interested in pragmatics, socio-linguistics, functional linguistics, discourse analysis and other

disciplines have proposed definitions that are useful in their work. Our purpose is to show the diversity and a cross-section of thought and contextualize our own approach to the concept of style. As with all the background material, much more could be included.

Table 1. Selected commentary and definitions of related terms

Term	commentary or definition	source
style	style is called "parole"	Saussure [Juc92]
style	style is "performance"	Chomsky [Juc92]
style	a "comparative concept" describing relative differences between "texts or discourses"	Jucker [Juc92]
style	Every single text has got a style as far as it has formal properties that can be compared with those from other texts.	Jucker [Juc92]
style	as "a selection of language habits the occasional linguistic idiosyncrasies which characterize an individual's uniqueness"	Crystal & Davey [Cry69]
style	A style is a consistent and distinguishable tendency to make some [of these] linguistic choices.	Karlgren [Kar04]
style	"style is the man himself"	George-Louis Leclerc de Buffon [Buf21]
style	"style is not an isolable quality of writing; it is writing itself"	Murry [Mur22]
style	"We no longer feel comfortable viewing style as the dress of discourse, the external and changeable garb for ideas…"	Walpole [Wal80]
style	style is "choice" or "option" and "not a unique and inseparable reflection of the author's (or student's) personality"	Walpole [Wal80]
stylistics	an "offshoot of linguistics" which "takes the position" that "style is a deviation from normal language usage"	Walpole [Wal80]
stylistics	successor to the ancient Greek concept of rhetoric	Biber [Bib89]
stylistics	"application of linguistics to the study of literature"	Fish [Fis81]
stylistics	"Stylistics is a method of textual interpretation in which primacy of place is assigned to language"	Simpson [Sim04]
stylometry	"an attempt to capture the essence of the style of a particular author…"	McEnery [McE00]
style checking	Style checking refers to checking text for errors that a book on good writing style, would discuss, such as	Heidorn [Hei00]

	overuse of the passive voice.	
stylistic meaning	those aspects of its meaning that are non-denotational, i.e., independent of the objects and events to which the text refers	Whitelaw and Argemon [Whi04]

2.3 A BRIEF HISTORY OF STYLISTICS IN 20TH CENTURY HUMANITIES

Richard Bradford in [Bra97] describes stylistics as a conceptual descendent of the ancient Greek *rhetoric*. During the early part of the 20th century, two different and disjointed groups were carrying on literary criticism focusing on what can be called "stylistics". The first was the Russian/Central European formalists. The other group was the British and American educators who practiced what later became known as "New Criticism." The two groups began cooperating during the 1960's and developing overlapping goals and methods. After the 60's, however, both groups had their academic predominance "unsettled" by a wave of interdisciplinary practices. Structuralism, post-structuralism, feminism and new historicism, all became significant elements of contemporary literary studies.

Bradford describes two new groups emerging in the last 70's called "textualists" and "contextualists." Textualists are comprised of New Critics and the European Formalists because "they regard the stylistics features of a particular literary text as productive of an empirical unity and completeness" [Bra97]. The contextualist group is comprised of the elements involved in the 60's movements such as post-structuralism, feminism and Marxist analysis. This type of stylistics involves "a far more loose and disparate collection of methods" [Bra97] but its members are unified in their concentration on the relation between text and context.

2.4 REACTIONS TO AND CRITICISMS OF STYLISTIC METHODS

Some commentators in a variety of disciplines do not consider stylistics as a legitimate or fruitful method of analysis. The main criticisms seem to be about the perceived arbitrariness of the meaning-assignment to stylistic methods and interpretation of algorithmically generated stylistic data. A certain stereotype of who stylisticians are and how they operate is often embedded in the criticism. Simpson summarizes the attitudes in this 2004 book:

> *There appears to be a belief in many literary critical circles that a stylistician is simply a dull old grammarian who spends rather too much time on such trivial pursuits as counting the nouns and verbs in literary texts.* ***Once counted those nouns and verbs for the basis of the stylistician's 'insight',*** *although this stylistic insight ultimately proves no more far-reaching than an insight reached by simply intuiting from the text."* [Sim04, page 2]

2.4.1 CRITICISM OF STYLISTICS' CLAIMS OF SCIENTIFIC OBJECTIVITY

Fish situates stylistics as "a reaction to the subjectivity and imprecision of literary studies" [Fis81]. He does not credit this reaction with adding any additional understanding, however.

> *The machinery of categorization and classification merely provides momentary pigeonholes for the constituents of a text, constituents that are then retrieved and reassembled into exactly the form they previously had. There is in short no gain in understanding: the procedure has been executed, but it hasn't gotten you anywhere. Stylistician, however, are determined to get somewhere…* [Fis81, page 55]

In addition, Freeman also attacks the profiling function, or emergence of stylistic categories, which is so central to stylistics.

> *The establishment of a syntax-personality or of any other paradigm is an* ***impossible goal,*** *which , because it is also an assumption, invalidates the*

> *procedures of a the stylisticians before they begin, dooming them to successes that are meaningless because they are so easy."* [Fis81, page 56]

Freeman reserves some strong criticism, what he calls the "heart of my quarrel with stylisticians," for the arbitrariness of what text features are or should be considered "significant." He says of the stylisticians:

> *…in their rush* ***to establish an inventory of fixed significances****, they bypass the activity in the course of which significances are, if only momentarily, fixed. I have said before that their procedures are arbitrary…* [Fis81, page 64]

For Fish stylisticians also commit the de-humanizing sin of creating meaning where none exists, at least not according to any existing literary theories. The new meaning assumptions become a circular justification for the utilization of stylistic methods.

> *The stylisticians, of course, have an alternate theory of meaning, and it is both the goal of and the authorization for, their procedures. In that theory,* ***meaning is located in the inventory of relationships they seek to specify, an inventory that exists independently of the activities of the producers and consumers, who are reduced either to selecting items from its storehouse of significances or to recognizing the items that have been selected****. As a theory, it is distinguished by what it does away with, and what it does away with are human beings, at least insofar as they are responsible for creating rather than simply exchanging meanings.* [Fre82, page 66].
>
> Fish also offers a psychological explanation for the stylisticians' behavior.
>
> *Behind their theory, which is reflected in their goal which authorizes their procedures, is* ***a desire and a fear: the desire to be relieved of the burden of interpretation by handling it over an algorithm****, and the fear of being left alone with the self-renewing and unquantifiable power of human signifying* [Fis81, page 66].

Simpson in [Car88] warns that stylistics analysis is too limiting since it has no access to the "extra-textual world of social, political, psychological, or historical forces" [Car88, page 7]. Carter and Simpson acknowledge stylistic trends to take such forces into account, but

> *…in spite the aforementioned sociolinguistic trends, most literary-stylistic analysis still sees referential, text-immanent language as a primary constituent of the text and as a locus of the author-initiated effects and response to those effects* [Car88].

2.4.2 STYLISTICS AS AN INESCAPABLY INTERPRETIVE ACTIVITY

The criticisms leveled at stylistics appear to concentrate on the subjective aspect of "interpretation" based on statistical data. One may be tempted to think that one could escape this criticism to a great extent in certain areas like computational linguistics, by simply not engaging in any interpretation of the results. However, [Car88] suggests this is not so easily achieved, if even possible.

> *The assignment of meaning or stylistic function to a formal category in the language remains an interpretive act and thus cannot transcend the individual human subject who originates the interpretation. Thus, while the recognition of specific formal features can in most cases by attested within the terms of the system,* ***the analyst has to be taken on trust in his or her interpretive assignment****. It is a perennial problem, or even dilemma, in stylistics that no reliable criteria can be generated whereby specific functions or effects can be unambiguously attributed to specific formal features of the language system.* [Car88, page 6]

This seems to suggest that even selection of language features to study and assigning of literal or figurative value to them could be considered problematic. But

what if one was to simply describe the findings without making any assignment to "specific formal features"? Even there, [Car88] says that one cannot escape "interpretation:"

> *Any resolution to describe the data rather than interpret it constitutes an interpretation. It is an interpretation of the way literary study can or should be approached and analysis of it conducted.* [Car88, page 6-7]

2.5 DEFENSE OF STYLISTICS

While having many skeptics in linguistics and literary criticism, there are those who have embraced and continue to advance the discipline of stylistics.

2.5.1 SIMPSON AND WALPOLE'S DEFENSE OF STYLISTICS

Simpson in [Sim04] answers Lecercle's attacks on stylistics.

> *Modern stylistics is positively flourishing witnessed in a proliferation of sub-disciplines where stylistic methods are enriched and enabled by theories of discourse, culture and society: feminist stylistics, cognitive stylistics and discourse stylistics are established branches of stylistics which have been sustained by insights from, respectively, feminist theory, cognitive psychology and discourse analysis.* [Sim04]

Simpson suggests that stylistic methods can best be defended when the "three R's" are observed: "Stylistic analysis should be rigorous, retrievable, replicable" [Sim04]. Walpole also offers redeeming values of stylistic methods and at the same time demystifying the statistical aspects for the liberal arts researchers. She quotes Corbett:

> *Some instructors try to heighten an awareness of style by a detailed scrutiny of published prose. As Edward Corbett suggests, "Any stylistic analysis must start out with some close observation of what actually appears on the printed page." This observation includes counting sentence lengths, types and numbers of syntactic structures, and classes and varieties of words-what DeQuincey called the "mechanology of style." Classroom applications of these analytic approaches can be found in professional journals. "Such a procedure would make counters and measurers of us all," says Corbett, admitting that this may seem repellent to humanistically-trained teachers; "but* ***this is a necessary step if we are to learn something about style in general and style in particular." For it gives us the requisite things to point to as we test the effect of authorial decisions****.*

2.5.2 DEFENSE OF STYLISTICS AS AN EMPIRICAL ACTIVITY

Giacomo Ferrari in [Fer03] justifies the move toward empirical and statistical methods because of the inadequacy of relevant linguistic theory. Ferrari offers the move toward computational linguistics as a natural result of linguistics' failures to offer better and more comprehensive models both in syntax and semantics. Regarding syntax, he writes:

> *Talking of syntactic approaches,* ***we observe that in many cases the move toward empirical studies is uncontrolled, because linguistics has no interpretation mechanisms****. Computational Linguistics produced important results in syntactic and morphological analysis because it could rely on robust theoretical linguistic models like Chomsky's grammar(s) and the theory of automata. This is not true in other fields like, for example, discourse and dialogue modeling.* [Fer03, page 163]

And here, he makes a similar criticism in semantics and suggests that linguistics theory should more fully embrace computational linguistics.

> *In the field of semantics, Computational Linguistics turned to logic because linguistics offered absolutely nothing for what concerns the meaning of sentences. Thus the conclusion is that Linguistics is in debt here, and should take advantage*

> *of stimuli coming from Computational Linguistics to* ***build new theories that take into account computational modeling****.* [Fer03, page 163]

According to some of the above researchers stylistics is a viable and growing discipline, but its models do not necessarily produce meaning or make claims beyond their own symbolic system. Stylistics thus need not be called a meaning-producing enterprise, but perhaps a hypothesis-checking enterprise validating or invalidating human theories.

Still a notion of style as a distinguishable "thing" is necessary in order to work with it. The problems with overlapping domains of style and meaning, or style and personality, remain.

2.6 Deriving a notion of style primitives: Style as "option"

In "Style as Option," [Wal80] Walpole tackles the aforementioned problem by stressing a conception of style as "choices among alternatives." She divides these choices into two groups: linguistic and extra-linguistic.

The framework represents perhaps the best operational definition for style that would allow robust models to achieve solutions to narrow problems involving style.

Walpole explains "What remains of prose after we have set aside the detachable ideas and the immutable requirements is Style, the vast area of writer's choice." She elaborates on the omnipresence of options in writing.

> *Options exist in small matters: the individual words, the optional comma that influences emphasis, the placement of a movable adverb, the different rhythmical effects of under and beneath. Options exist in sentence variations: coordination vs. subordination, clausal vs. phrasal constructions, polysyndeton vs. asyndeton. Options exist in larger decisions: whether to use parallel sentences, whether to explain a point in depth or superficially, whether to illustrate a concept with several short examples or one long analogy. And options exist in the total work: the level of diction, the attitude toward the audience, the weight given to flourishes or simplicity. Though* ***writers may be constrained by imposed limitations, they have in each of these areas wide freedom of selection****. And in each of these areas they produce, in Ross Winterowd's phrase, "****features that we can 'point to.'****"* [Wal80]

These features we "can point to" are perfect candidates for "style markers" that we can use to assess the presence of that style.

2.6.1 TWO ASSUMPTIONS IN WALPOLE'S NOTION OF STYLE

Walpole also explains that style as "option" necessitates two assumptions. First, that the style is actually a conscious choice by the author. The second assumption is that style is a variable quality not necessarily always present with the same intensity and "not a unique and inseparable reflection of the author's (or student's) personality" [Wal80].

This view of style accommodates the everyday observation which is that writers can and often do deliberately change their own style. A view of style "inseparable" from writing would not accommodate this. At the same time, a component of these choices may be unconscious. This, we think, Walpole would call an extension of personality rather than style.

Walpole's own background in pedagogy allows for more unique observations allowing us to entertain the possibility of altering and transforming written style.

Literature students are already doing a form of stylistic transformation exercise, as Walpole observes:

> *A second approach sometimes followed involves exercises of* ***deliberate imitation****. Corbett again provides guidance in this traditional technique. Students learn how to* ***choose words and structures that echo the choices of an acknowledged style master****. As they clothe their own matter in a borrowed manner, they are sensitized to the range and impact of writing options* [Wal80].

She describes another exercise where students are encouraged to highlight and later subdue "obvious style" [Wal80].

2.7 Related work on style markers

Style markers, also referred to as "features" [Cry69][Bib88] or "discriminators" [McE00] are indications or consequences of choices that were made in writing in a recognizable way. Uniform and extensible markers, with well-defined scopes are necessary for an automatic style analyzer. One objective of our project is to publish a list of such markers. We aim for this list to be as comprehensive as possible. We surveyed the literature in order to compile a preliminary list of existing markers used for various related purposes in stylistics.

Various works have offered examples of, as well as process suggestions for working with style markers. Mendenhall studied word length distributions as

systematic style markers as far back as 1887 [Men87]. Yule studied sentence length in prose [Yul38] and vocabulary statistics [Yul44] as distinguishing features. Crystal and Davy [Cry69] informally identified some stylistic features that they considered important and even devised measures of comparison for them. DeQuincey's notion of "mechanology of style," includes "counting sentence lengths, types and numbers of syntactic structures, and classes and varieties of words" [Wal80].

2.7.1 MARKER WORDS AND DISCRIMINATORS

Mosteller and Wallace's [Mos84] classic study of the *The Federalist Papers* relied on specific "marker words" which were a series of 30 mostly function words like "upon, whilst, there, on, while, vigor, by, consequently." [Tay87] also used specific functions words in their study of Shakespearian works. Ten words used in that work were: *but, by, for, no not, so that, the*, *to* and *with*. Merriam in [Mer93] used Taylor's 10 words in her own Shakespeare/Marlowe study using standard scores, and later in [Mer96] added more combinations to the original 10. *The Federalist Papers* were also studied by Holmes and Forsyth [Hol95] using six distinct vocabulary measures.

McEnery and Oakes [McE00] describe five classes of "discriminators" in their basic methodology. These are ranges of "features that successfully discriminate among the authors" [McE00]. The five classes are:

1. Word or Sentence Length (such as average number of words per sentence)
2. Vocabulary Richness (such as *Simpson's rule*, *Yule's K, Honore's R* and entropy methods)

3. Word Ratios
4. Letter Based (frequencies)
5. Part of Speech

2.7.2 SYNTACTIC RULES AND GRAMMAR BASED MARKERS

While "most studies are based on word forms and frequencies in occurrence" [Luy04] or are "lexical in nature" [McE00], mainly due to ease of computation and lack of large corpora with sophisticated annotation, there has been an increasing number of studies that are venturing beyond simple word counts and vocabulary metrics. Baayen et al. used syntactic rewrite rules, such as verb phrase (VP) or noun phrase (NP) production rules, as primary markers for authorship attribution [Baa96]. The study found that syntactic marker frequencies are at least as good as word frequencies in authorship attribution predictions [Baa96]. Similarly, [Arg98] used part of speech n-grams in their study and found them effective as markers.

Heidorn in [Hei00] briefly discusses the style-checker component of Microsoft Word 97. That commercial product uses a set of 21 or so "grammar and style options," each of which is a rule that checks for a specific stylistic phenomenon, for example "correct capitalization," and "misused words." Microsoft has coded these rules developed by their team of linguists.

The Microsoft notion of "style" here is a single two-dimensional discrete variable. At one end, "formal", all 21 options are checked for. At the other end, "casual," none is checked for. The MS-Word 2002 product, a successor to the one discussed in

[Hei00] uses 22 distinct options for style. Many like clichés, contractions, gender-specific words, use of first person, numbers, sentences beginning with "And," "But," and "Hopefully" use pre defined assets that they check for. Others such as "wordiness", "sentence length", "hyphenated and compound words" and "successive nouns/propositional phrases" do procedural checking using NLP tools.

Lyuckx and Daelemans identified four types of features in [Luy04]:

1. *Token-level* such as word lengths, syllables and n-grams
2. *Syntax based* like part-of-speech and rewrite rules
3. *Vocabulary richness*: type-token ratio, hapax legomena
4. *Common word frequencies*

An interesting example of (2) from above may be the Brill Tagger [Bri04]. Eric Brill and his team at Microsoft developed a two tier system whereby a relatively dumb part-of-speech (POS) tagger initially tags all words based on their statistically known most likely POS. Subsequent routines check for specific patterns and alter the tags if they find them. The routine uses variables that stand in for words three lexemes ahead and three behind the current word. Rules are specified in terms of these variables and the current word-tag tuple being considered for alteration [Bri04].

2.7.3 NONLINGUISTIC AND STRUCTURAL FEATURES

De Vel in [DeV01] explored structural features: Non-linguistic text layout or meta-data such as titles, author or paragraph position. The problem with these features

is that they are not standardized and not every context has them available. Thus while online blog posts may have timestamps and category selection, these are lacking in short stories. Structural features were also used in [Zhe06] and [Abb08].

2.7.4 WORKING WITH LARGE NUMBERS OF FEATURES

According to [Arg03], "stylometric models have been based on hand-selected sets of content-independent, lexical, syntactic or complexity-based features." In [Juc02] the authors discuss how to proceed after a given super-set of style markers is provided. Three specific points are especially of note. A "stylistic analysis" phase should concentrate on the best features among all that is available based on which ones exaggerate the difference between the two texts being compared. Second, any frequencies have to be seen in relation to the length of the text so that "we should talk in terms of density, that is to say the frequencies of a feature within a well-defined stretch of text" [Juc02]. Lastly, the authors repeat a point from Enkvist in 1980 suggesting that the "guiding principle" in comparison of feature sets should be to keep as many non-linguistic features as possible constant over all the texts to be compared in order to be able to assign the linguistic difference with more confidence to those few features that do vary [Juc02].

Latent Semantic Analysis [Lan98] and Probabilistic Latent Semantic Analysis [Lat08] are techniques used to determine word relationships. In these techniques a larger set of initial distinct tokens may be reduced as more words are found to be equivalent on semantic grounds.

Currently many writing analysis techniques are used in NLP, Computational Linguistics and readability assessment. WordNet [Mil90], an online word ontology tool, can be used to calculate statistics of different word groupings. The Gunning fog index [Haa07], and the Flesch-Kincaid readability test [Haa07] are two popular means of assigning reading level by analyzing vocabulary in writing.

Authorship attribution works from Khosmood and Levinson [Kho05][Kho06] contain an enumerated set of criteria used as features. Their markers include vocabulary measurements [Fak01], lexeme statistics [Kes03], as well as semantic constructions.

Works of Abbasi and Chen have consistently used high number of markers based on previous use in literature. The applications were text attribution in asynchronous conversation (Web forums and email) with hundreds of documents. In [Abb05], 418 features were used and in [Abb06], 106 were used. The system described in [Abb08] used at least 327, but possibly hundreds more were considered before a feature selection phase would reduce the number.

Both Abbasi and Chen [Abb08] and Zheng et al. [Zhe06] provide enumerated feature breakdowns of their markers, as well as categorization for them. In [Zhe06], the authors placed 270 features into four categories: *Lexical* (further subdivided into *Character-based* and *Word-based*), *Syntactic, Structural* and *Content-specific*. In [Abb08], the authors present almost exactly the same breakdown with only one high level category, *Idiosyncratic*, having been added to capture identity based misspelling trends.

In their survey paper, [Ste09] present a more comprehensive multi-layer categorization of features that captures much of the previous literature on the subject.

Table 2. Types of stylometric features as presented in [Ste09] survey paper

Categories	Features
Lexical	Token-based Vocabulary richness Word frequencies Word n-grams Errors
Character	Character types n-grams (fixed and variable length) compression methods
Syntactic	Part-of-speech (POS) Chunks Sentence and phrase structure Rewrite rule frequencies Errors
Semantic	Synonyms Semantic dependencies
Application-specific	Functional Structural Content-specific Language-specific

2.8 Related work in automatic authorship attribution systems

There have been a few systems that we have encountered which have attempted to bring together style marker analysis, feature extraction, feature selection and finally authorship attribution or similarity detection as their main function. Of those, two are of note and particularly relevant to our work. These are JGAAP [Juo09] and the Writeprints system [Abb08]. More discussion on these systems will be followed in chapter 4.0 where we learn from these systems and compare their features with our own contributions.

2.8.1 JGAAP

JGAAP [Juo09] is a system developed by Juola et al. for authorship attribution. JGAAP was developed after and response to the 2004 Ad-hoc Authorship Attribution Contest (AAAC) which was also organized by Juola. The software is autonomous and user friendly. It features a "wizard" like interface whereby the user selects documents and labels them, applies pre-processing operations, chooses event sets (features) and distance measures all on separate consecutive screens. The program then computes an attribution matrix associating unlabeled documents with one of the classes that had labeled documents. It is simple and user friendly.

Juola et al. have pre-loaded the system with both corpora from the AAAC and a selection of popular features and distance formulas they encountered in the contest. This makes it easy to reproduce results from the AAAC as well as repeat the same mythology on other corpora.

The event sets and distance measures are somewhat limited in their number and variety, but the system is open source and modifications can be made to it, complete with new formulas and analysis capabilities. Probably the biggest weakness that we see with JGAAP in its current form is its inability to consider multi-marker ensemble methods and large feature sets. The designers appear to have a preference for dynamic event sets which are not universal by definition and problem dependent and consequently, also difficult to combine.

We use JGAAP in the present study, both to augment our list of markers and also to run attribution experiments and analysis.

2.8.2 THE WRITEPRINTS SYSTEM

Writeprints is the core approach to a system described in [Abb08], with background concepts discussed in [Li06] and [Zhe06]. It is capable of large-scale analysis and identity-level attribution. The system extracts and uses a much larger feature set, "327 lexical, syntactic, structural and content-specific features," [Abb08] which they call "Baseline Feature Set" (BF), and an "Extended Feature Set" (EF) which are comprised of n-gram (words, characters, Parts of Speech and digits), as well as idiosyncratic markers such as common misspellings. However, a feature selection phase, based mainly on Information Gain (IG) is applied to the features and reduces their number significantly for the final classification operations.

We were not able to download and execute the Writeprints system itself and report results based on [Abb08].

2.9 RELATED WORK IN PARAPHRASE GENERATION

Automatic paraphrase generation is a concept that can be approached from a variety of different perspectives including Machine Translation (MT) or Natural Language Generation (NLG) or Statistical Natural Language Generation (SNLG).

A paraphrase does not necessarily mean a synonym. Many linguists consider a paraphrase as derivative text that retains "approximate conceptual equivalence" with respect to the original [Bar01].

We can broadly divide the related works in paraphrase NLG into "concept-to-text" and "text-to-text" systems. In this thesis, we are generally pursuing a text-to-text approach. However some important work in concept-to-text has been done which we will cover briefly below.

2.9.1 Related work in concept-to-text paraphrase generation

Concept-to-text generation systems take their input in form of machine representation and produce natural language text or utterances. Often these systems are part of a full NLU-NLG cycle whereby the ultimate input is actually text, but there is a crucial and necessary stage in the middle where the information content is stored as machine representation and machine operations could modify this representation.

Concept-to-text NLG systems tend to have much greater domain coverage and consequently much wider range. Such NLG systems have an advantage in working with fully "understood" concepts and handcrafted realization rules. Because they typically interface with surface text realization system, this allows them to generate output language with much greater accuracy and control. In much of the NLG literature, the goal is to generate language for specific informational purposes. Starting perhaps with [Hov88] researchers have begun experimenting with generating language not only for its informational content, but also for its style, emotion, or manner of delivery. Insofar

as concept-to-text NLG systems can generate closely related sentences differing only in some defined dimension of style, they can be considered paraphrase generators.

In [Ink06] the authors present a novel plan to alter the semantic orientation of a passage with automatic sentence regeneration of sentences where strategic near-synonyms exist. The goal was to replace certain emotion words with near-synonyms exhibiting more favorable or less favorable tone. The authors used standard NLP techniques (POS tagging and WSD) to find candidate sentence patterns containing desired emotion words, populate the variables of an "Interlingual representation" object, then perform the correct substitution. The final step, common to all concept-to-text approaches, was to use a surface text realizer to output the final sentence. The realization tool was Xenon [Ink03], itself based on statistical realization tool HALogen [Lan02][Lan94]. The results from [Ink06] were inconclusive. Human judges could not substantially agree on whether the passage had moved to a more positive or less positive direction.

More recently, researchers working with PERSONAGE [Mai08] generated personality-variant paraphrases with a high degree of control. PERSONAGE was designed as a dialogue utterance generator with a sophisticated NLG pipeline [Wal02] including a full featured surface realization tool: RealPro [Lav97]. Based on "Big 5" personality theory, PERSONAGE is a highly parameterized generator of recommendation and comparison commentary about New York City restaurants. In

[Mai10], human judges successfully recognized generated utterances along two psychological dimensions of *Extraversion* and *Emotional Stability*.

2.9.2 Related work in text-to-text paraphrase generation

The "automatic" part of automatic paraphrase generation means that equivalence relations are either encoded in linguistic transformations or are learned from external sources; thus never directly from human experts. The external sources can be *parallel* corpora (direct restatements such as different translations of the same novel) or *comparable* corpora (same content, such as different reports of the same news event) or lexical databases such as thesauri [Bar01].

The first attempt to learn significant number of rules from corpus was [Jac97] where the authors successfully derived synonyms for many technical terms. A pioneering work using multiple English translations of the same foreign language source was done by [Bar01]. The authors used self-labeled parallel sentences to derive morpho-syntactic rules for transformations that they used to create examples to train a classifier. The classifier worked with about 9,500 sentences. A random selection of 500 sentences was evaluated by human judges and they found a high degree of overlap with their automated system. In [Kes10], the authors claimed improvement over [Bar01] by being able to extract paraphrase emotion terms from text without requiring a parallel corpus or word alignment. They used a limited number of "seed terms" to bootstrap on and generate a large number of contextual patterns matching paraphrase candidate sentences.

2.9.2.1 Siddharthan's simplification algorithm

Advaith Siddharthan in [Sid10] used high-level pre-defined syntactic rules that operate on typed-dependency structures (a dependency tree representation of sentences). Each rule could potentially modify one such structure corresponding to one sentence, but rather than to reestablish a parse tree and realize the sentence afterwards, the authors de-parse the (modified) tree and use the old structure to reconstitute a new sentence. This "shallow method" has shown great promise in [Sid10] where authors tested it on some token meta-rules and over 1100 sentences. They found high precision and low recall with this scheme.

We use a version of the system in [Sid10] designed for sentence simplification as a transform in our system (see Section 5.2.6).

[Zha09] introduced Statistical Paraphrase Generation (SPG) where paraphrases are generated for a specific measurable need. In [Zhu10] the authors experimented with fitting statistically generated paraphrases to specific reader profiles.

A general problem with substitutability of paraphrased sentences has been identified by the community. Not every context allows every paraphrased sentence to be seamlessly substituted for the original [Zha07]. This further underlines a need to independently validate substitutable material for the target context. Both [Cos11] and [Bou11] used the Internet as a validation tool. In [Bou11], the authors specifically experimented with validating sub-sentential French-language paraphrases in context. They generated paraphrases from multiple sources including Wikipedia edits, human

crowd-sourcing (via a game), and translation by pivot (translation into Spanish or Chinese, and back into French) and chose among them using a formula involving web search engine hit ratios of the paraphrased to the original.

The results were compared to two baselines: one based on simple web-counts, the other based on syntactic dependencies. Evaluated by two judges, the results indicated the [Bou11] system outperformed simple web-count baseline by 2-3%, and dependency based baseline by about 20%. The authors also concluded that Wikipedia based paraphrases were far more likely to be considered a good substitute than other acquisition sources.

Using the Moses statistical machine translation toolkit to generate paraphrases, the authors in [Wub11] compared a phrase based with a syntax-based approach. Phrase-based approaches compare chunked sections of a sentence as a unit and replace one or more chunks to produce a paraphrase, while syntax based produces a parse tree and replaces the entire tree to produce a paraphrase. The authors used a 2.1Million token corpus in Dutch to learn paraphrases. They also presented a novel evaluation measure based on the NIST [NIS02] MT metric and the Levenshtein [Lev66] distance, to maximize quality with bias toward the most dissimilar paraphrase.

2.10 Related work in style obfuscation and imitation

Style obfuscation is loosely defined as deliberate changes to one's writing in order to hide the true style of the text that is presumably attributable to the writer. Style *shifting* is "automatically adjusting from one style to another," [Tse04]. When that

shifting is performed with a target in mind, it is considered style *imitation*. The applications that have been discussed for such operations range from preserving anonymity to digital forensics. The works of [Kac06] [Bre09] and [Juo10], are more specifically concerned with exposing weaknesses of authorship attribution systems.

2.10.1 LANGUAGE OBFUSCATION TO PRESERVE ANONYMITY

In [Rao00], the authors approached the issue from the standpoint of online anonymity. They find just by using a few syntactic and semantic markers, one can easily identify a large percentage of anonymous newsgroup and email authors (given enough exposure into their true writing). The "countermeasures" section offers a few suggestions on how to circumvent this kind of identification. They advise:

> *It should be easy to design minimally disruptive countermeasures to tackle most syntactic leakage. However, fixing the problem of vocabulary requires more investigation. For instance* ***a thesaurus tool could prompt the user to consider alternatives*** *while composing messages, thereby reducing variations in vocabulary.*
>
> *However, countermeasures to address semantic leakage appear to be hard to design. One drastic (and somewhat facetious) approach is* ***to use natural language translation systems to translate a document from English to another language and back.*** *While this may destroy the function word frequencies in the original document, it would considerably change the meaning of the message, given the primitive state of translation systems. Even this would not destroy information about the meta-level concepts that an author uses in his documents.*[Rao00]

We note that the source of their last reservation, expressed in the final sentence above is that "meta-level" concepts would not be altered. This is because a significant amount of the semantic material used in [Rao00] is related to topic and framework of

the content, which tends to correlate to individual author in most online settings, but only coincidently. Thus the point made is perfectly valid from an individual anonymity point of view that is in fact the approach of the authors. But from a stylistic point of view, "meta-level" concepts can be expressed in multiple styles. We consider content to be irrelevant in true "stylistic" obfuscation.

In "Obfuscating Document Stylometry to Preserve Author Anonymity"[Kac06], Microsoft researchers discuss what characteristics a modified document would have to have to invalidate the relevant authorship attribution techniques. One of the questions they specifically pose is "How resilient are the current authorship detection techniques to obfuscation?" The authors use a standard corpus (*The Federalist Papers*), with word frequencies as features and linear kernel SVMs for attribution. The authors were able to generate interesting statistics on specific word usage that would have to be increased or decreased in order to accomplish the obfuscation goal, but they made no attempt to actually induce those changes. Consequently they did not consider the consequences of replacement procedures and the entailing language changes in their experiments. The impact of introducing "replacement" language, in order to achieve a real obfuscation case, was acknowledged but not quantified. The authors' conclusion theorizes about future possibilities:

> *Given this result, it is not unreasonable to expect that a tool could be created to provide feedback to an author who desires to publish a document anonymously.* ***A sophisticated paraphrase tool could theoretically use the function word change information to suggest rewrites that worked toward the desired term frequency*** *in the document.*[Kac06]

2.10.2 OBFUSCATION AS A CHALLENGE TO AUTHORSHIP ATTRIBUTION SYSTEMS

Brennan and Greenstadt in [Bre09], use a different approach to boundary-test authorship attribution systems against what they called *obfuscation attacks* and *imitation attacks*. The authors conducted a user study whereby users were asked to provide existing sample writing material, and also obfuscate their own writing style in an additional short passage they were instructed to produce. In addition, they also had to write a second passage in which they try to imitate a specific style (that of Cormac McCarthy's *The Road*). The results showed that both obfuscation and imitation attacks were highly successful, causing their authorship classification systems to perform significantly below chance on the "attack" documents in most cases. The same systems would perform very well on the original sample writings without any obfuscation or imitation. Juola in [Juo10] duplicated the basic problem on the same corpus (12 of the 15 authors) using JGAAP [Juo09]. 160 experiments were conducted using JGAAP with various event sets and canonization parameters. While the results were not quite as conclusive as [Bre09], Juola writes that "no method was able to perform 'significantly' above chance…" on the corpus [Juo10].

"Obfuscation attacks" undermine authorship attribution claims made by researchers who accept the "homogeneity of authorship" assumption [McE00], meaning that there is a detectable trace of the "author" in every text he/she is written, even the ones where the author is deliberately altering the writing. The assumption suggests there exists an invisible watermark for each author. However, McEnery and Oaks

dispute this and write that "there is no clear undisputable evidence of the existence of such features" [McE00]. They use the example of the notoriously versatile Anglo-Irish writer Oliver Goldsmith, whose anonymous works could be easily attributed to multiple authors [McE00].

2.11 Our approach to style

It is important to discuss the debates in the community (as done above) regarding the very feasibility of stylistic analysis. Specifically, a foundational assumption of this work is that style and meaning are separable and can be processed computationally. Thus we adopt the Walpole definition of "style as choice".

From the various communities doing relevant research in the area that we presented in this chapter, our approach draws most closely from two groups: The authorship attribution community (section 2.8), including the recent discussions around obfuscation, and the text-to-text translation community (section 2.9). From the latter we adopt transformation techniques targeted toward stylistic changes.

3.0 MARKERS AND ANALYSIS

We view style markers as the building blocks of style. In this chapter, we outline how we come to **collect, organize and evaluate style markers**. Our model of style is dynamic and dualistic with respect to meaning. An author's choices and distinctiveness of the writing can be measured by a statistical analysis of writing features or markers. This applies not just to individual writers, but it is valid for any sources of writing material exhibiting a distinct style.

Our first obstacle is to identify and isolate markers that are relevant to style. This problem itself remains unsolved, thus our approach is one of encoding and enumeration of a large set of possible stylistic markers that are performatively evaluated and weighed by experimentation. While we cannot say that we will have discovered a definitive set of all style markers, we can evaluate the ones we have and experimentally show vectors of markers that perform well in actual style-based classification exercises.

3.1 WHAT ARE STYLE MARKERS?

There are some notions of style in linguistics and corpus studies. However, none are comprehensive and quantifiable enough for our work. As previously stated, we

intend to derive the relative importance of markers based on actual performance in experimentation. However, we do need a superset of all markers to evaluate.

To gather an initial set of markers, we survey the literature for markers used in actual experiments in the past (See Section 2.7). However, text segmentation and tokenization [Pal00] processes are not standard. In addition, there are many non-standardized ways in which the same concepts have been used as markers, and many possibilities for parameterization exist. Even if the we find ourselves having to first standardize the entire concept in a comprehensive marker taxonomy. Having achieved the structure for such a taxonomy, we could then intelligently place all markers that we find in the literature into a definitive unambiguous hierarchy.

3.2 MARKER TAXONOMY

A systematic categorization of style markers allows for more accurate evaluation and comparisons. In this paper we present a set of style markers, and propose a hierarchical extensible taxonomy for them.

We examine the literature, including the AAAC submissions where researchers have experimented with various recipes. JGAAP [Juo09] contains a collection of many of the style markers used by participants of the AAAC. We augment the markers in JGAAP with other markers from a variety of sources and we add some statistical measures to create a standardized set.

3.3 DOCUMENT MODEL

The fundamental assumption in any feature extraction is that the extracted data is a representative model of the underlying text corpus. In computational stylistics, we are making the same claim that "style" can be modeled by various markers and associated statistics. In order to accomplish the necessary evaluations, we must ensure that markers remain as universal as possible. Thus we minimize introduction of markers that may not be applicable to some texts, or may be direct reflections of the size of a corpus making comparison against a corpus of a different size meaningless. For example we would like to compare the style as exhibited in a single page against that of an entire book.

In general, we hold the following assumptions about the corpora we are considering for:

1. Each corpus is a collection of documents or just one document.
2. Each document is divided into paragraphs.
3. Each paragraph is divided into sentences (not necessarily a well-formed linguistic definition of sentence).
4. Each sentence consists of words.
5. Each word consists of characters.

Our standard object of comparison is the corpus. Since a corpus can be a single document for our purposes, we can compare one chapter, one page or even one

paragraph (corpus of a single document, single paragraph) against books and collected works.

Lexical tokenization routines have traditionally been necessary in the pre-processing phase [Pal00]. However, in order to support our goal of unbiased flexibility, we have combined such pre-processing phase operations with the individual markers as parameters.

3.4 TAXONOMIC HIERARCHY

We propose a taxonomic hierarchy. It consists of the following elements:

1. Categories: These are high-level functional themes for the marker collections. Examples: *lengths, words, N-grams, readability measures*
2. Families: Families are middle level categories describing the type of markers in the concept. Examples: *characters, words, sentences*
3. Markers: Markers are base level stylistic features whose presence we are measuring. Examples: *words per sentence (wps), words per paragraph (wpp), words per document(wpd)*
4. Parameters: Parameters could be thought of as sub specification of markers, i.e. variation of markers that may be used simultaneously. In most techniques the parameters correspond to canonizers or pre-processing phase routines that operate on the entire corpus, resulting in the same "parameters" for every marker, as the corpus is modified prior to extraction of any features. This is the

way JGAAP handles them and it functions efficiently since only one event set (feature) is being considered at a time. But for ensemble methods, we use multiple markers with different parameters each. Examples: *Unify capitalization, unify numerics (replace with token), exclude common words in language X*

5. Statistics (or summary statistics): These are ways of summarizing marker output for the purposes of comparison, classification and machine learning. While concepts, families, markers and parameters mainly describe a feature event, quantifiers specify how to represent data from the said events. Statistics control the footprint of each marker in the final feature matrix. For example a summary statistics function for the marker "characters per word" (list of all word lengths in a corpus) could return any of the following:
    a. A large array of word lengths over the entire corpus (null quantifier)
    b. A single value, the average of all word lengths
    c. A 5-tuple of floats denoting minimum, maximum, mean, median and variance for the array.

The hierarchy can be represented as an acyclic graph whose leaves are *marker instances* that become part of the style representation of a corpus.

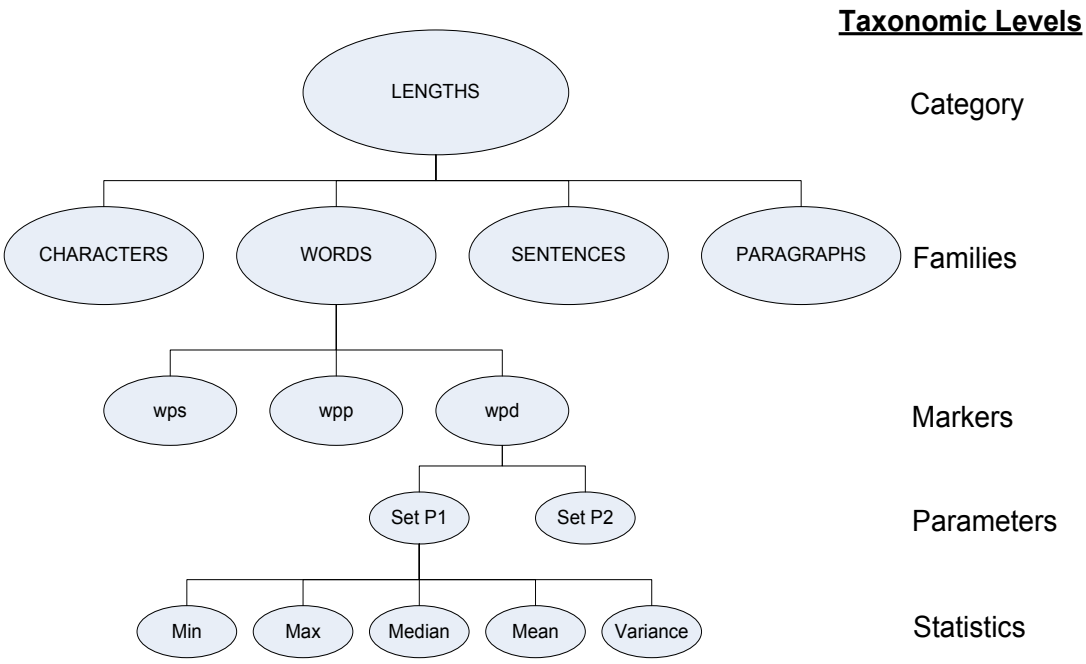

Figure 3. Partially specified taxonomy for *Lengths*

3.5 TYPES OF CATEGORIES

Categories are the highest-level distinctions of the taxonomy. We outline the implemented categories [Kho11] and the families associated with each. Neither the categories nor any of their sub-structure is meant to be a permanent statement on marker style markers or their organization.

Table 3. Implemented categories from [Kho11]

Category	Description	Families
Lengths	Counts and sizes of text features such as sentences and paragraphs.	Characters, words, sentences, paragraphs, syllables, numerics, vowels, punctuation, symbols
Words	Counts of many categories of words, word frequencies and word lists.	Most frequent words, least frequent words, parts of speech, misspellings, word lists (dictionaries)
N-grams	Counts sequence-based counts of characters, words and parts-of-speech.	Characters, words, parts of speech
Readability	Measures presence of well-known readability, sentence complexity and phrases recognized as cliché or poor communication.	Readability (Coleman-Liau, Kinkaid, Flesch Reading, Fog index, ARI, Lix) and Complexity (syntactic depth, parser phrase counts), GNU Diction rules (cliché, rewrite, run-on sentence, superfluous language)
Semantics	Measures having to do with meaning and word senses.	Word Net synset size, synset depth and distance

3.6 MARKER EVALUATION

One of the principle uses for having a comprehensive taxonomy is that it allows us to easily evaluate subsets of markers against particular style classification problems. In this section, we show and compare marker performance across multiple problems, and observe marker performance. Of course any conclusion on an experiment like this is directly reflective of the corpora that it was evaluated on. For best results, the conclusions should not be depending on peculiarities of a certain corpus, for example

size, genre, medium, etc. Instead we should aim to experiment on the most diversified set of texts as possible.

We report on experiments conducted in [Kho11] where 502 marker instances were evaluated using the AAAC corpus [Juo09]. We have reproduced a subset of the relevant markers from [Kho11] in Appendix B of this document. In Section 3.6.1 below, we briefly describe the AAAC corpus.

3.6.1 THE AAAC CORPUS

The Ad-hoc Authorship Attribution Contest (AAAC) was held in 2004 with many researchers participating [Juo04]. The AAAC corpus is particularly well suited for our marker evaluation task for several reasons. The most important reason is that the corpus has purposefully provided significant level of diversity in its many problems.

Documents of different problems differ with each other in genres and text types, as well as document sizes and training/test size ratios, but remain highly uniform within each problem. The corpus is created and prepared by contest organizers and is available to anyone in its original form, allowing for ease of repeatability. The formatting is machine friendly and in fact, has already been used as an example corpus bundled with the software package JGAAP [Juo09]. The texts are used exactly as distributed in plain text files without the need for any further preparation.

The AAAC corpus is divided into multiple problems. Each problem consists of a set of unlabeled test documents and a set of labeled documents associated with an

author. Usually multiple labeled documents exist per author that can serve collectively as a training corpus, allowing for cross-validation. However, problem H only has one training document and one test document for each author.

The types of writing in the problems themselves are diverse. They include student short essays in American English (problems A and B), novels (problem C and G), plays (problems D and E), letters (problem F) and speech transcripts (problem H). Problems I through M are in French, Serbian-Slavonic, Latin and Dutch respectively and are not used in our study mainly because many of the markers used are English specific.

Participants in the AAAC utilized many algorithms each depending on a relatively small set of features extracted from the contest texts [Juo04][Juo06][Juo09]. Each algorithm/feature set/parameter set can be thought of as a "recipe" for authorship attribution. The compositions of the recipes, as well as the procedure to apply them, were entirely the work of individual participants based on their own hypotheses.

For this experiment to evaluate markers, we use the training sets of the eight English language problems from the AAAC corpus: problems A, B, C, D, E, F, G and H. This represents over 50 authors and 187 total documents with known authors.

3.6.2 MARKER EVALUATION METHODOLOGY

We extract 502 marker instances from all training document collections [Kho11]. This results in a single set of markers for all same-labeled documents within a problem, as well as individual sets for each document. These 502 cover most of the

taxonomic markers we presented above with the notable exception of sentence-level complexity statistics (phrase count and syntactic depth).

We use all of the "N-gram" markers with N <=5 and number of top n-grams = 300 for characters, words and parts of speech, meaning the statistics are derived only from the top 300 n-gram events. There is no unification of white space, elimination of numbers and symbols or capitalization for n-gram based markers.

For all word counts, we use alphabetic words only, although for sentence and paragraph lengths, we used all space-separated tokens. For dictionary based operations (stop words, top 1000 English words and top 100 non-stop words) we use uniform capitalization.

For all array-based markers, we use standard *statistics* that reduced all arrays to a 5-tuple of minimum, maximum, mean, median and variance. For single value markers (such as *xLegomena, readability indices*) as well as for top 100 non-stop-word frequencies, we do not use any summary statistics, only the raw frequencies.

We use Python with NLTK tools [NLT09] to extract most of the markers. Part of speech tagging is done by the Stanford Tagger [Tou03], and the Link Grammar parser [Gri95] is used for some phrase level semantic statistics.

3.6.3 MARKER EVALUATION RESULTS

We consider each of the 502 marker instances un-weighted, in isolation. We use the Euclidean distance for comparisons and perform attribution on the problem set based on finding the labeled corpus with the minimum distance.

The eight AAAC problems (A-H) have 187 labeled documents total representing the work of over 50 authors. We classify each document with the given choices in the corresponding problem. For each marker we count the total number of documents (out of 187) that were classified correctly. This is the "absolute" attribution performance of a given marker, given in the following formula:

$$PERF_{m,absolute} = \frac{\sum_{i=1}^{NUM\_PROBLEMS} C_{i,m}}{\sum_{i=1}^{NUM\_PROBLEMS} N_i}$$

where m is a marker-instance, N_i is the total number of documents in problem i, $C_{i,m}$ is the number of documents correctly attributed by marker m and *NUM_PROBLEMS* is the total number of different problems, or 8 in this case.

If we only have absolute performance, then a problem that has a large number of documents to be classified could really dominate the evaluation of the markers. Thus, we also calculate the results in terms of problem-relative and marker-relative correct attributions.

The problem-relative results (called adjusted relative performance) are displayed in Fig. 4, along with the absolute performance for each marker. The Problem-

relative performance numbers are calculated by considering the percentage of correct attributions per problem, regardless of the how many documents it has. The following formula provides the relative performance of marker *m*.

$$PERF_{m,relative} = \frac{\sum_{i=1}^{NUM\_PROBLEMS} \frac{C_{i,m}}{N_i}}{NUM\_PROBLEMS}$$

For example if a problem had only 2 training documents, and only one of them was correctly attributed by marker X, then it would have 50% problem-relative correct attribution for X. If the problem had 10 training documents and still only one was correctly attributed, then X would have 10% attribution score. If one marker was able to correctly attribute every document in the problem it would achieve 100% problem-relative score. Thus, the relative value eliminates the size of the problem as a factor.

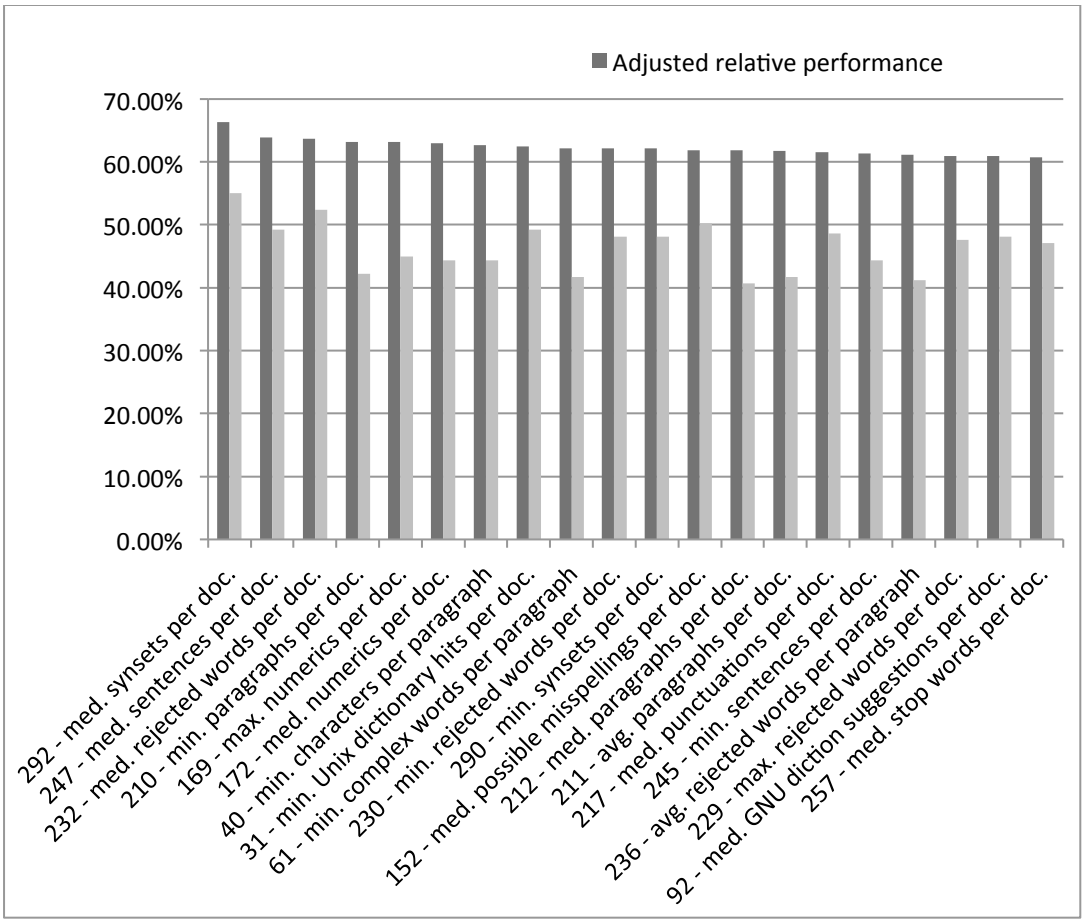

Figure 4. Best performing marker instances by relative correct attributions (absolute correct attributions also displayed)

Every attribution problem does not have the same degree of difficulty. We are also interested in marker performance relative to other markers per problem. To accomplish this, we rank all the markers from the best performing to the worst for each problem attribution, and calculate difference from the mean in units of standard deviations (z-scores) for each marker, for each problem.

Table 4. Marker instance performance relative to other marker instances (see Appendix B for Marker codes)

marker code	average z-score	relative performance	absolute performance
292	1.65467884	66.29%	55.08%
232	1.47054735	63.72%	52.41%
247	1.34549143	63.88%	49.20%
152	1.29048525	61.84%	50.27%
310	1.24059087	62.47%	49.20%
229	1.22346475	60.98%	47.59%
40	1.21858176	62.66%	44.39%
230	1.21157427	62.12%	48.13%
290	1.21157427	62.12%	48.13%
217	1.21012445	61.59%	48.66%
169	1.18920219	63.19%	44.92%
14	1.17039029	59.15%	44.92%
172	1.15087611	62.98%	44.39%
92	1.13307073	60.92%	48.13%
8	1.12271391	59.68%	45.45%
44	1.11259672	60.44%	45.99%
257	1.11227455	60.73%	47.06%
289	1.11002161	59.69%	48.66%
245	1.10662925	61.31%	44.39%
234	1.08100501	59.34%	47.59%

3.6.4 MARKER EVALUATION ANALYSIS

In Section 1.4.1, we described an "analysis" function of our system, whereby users can analyze a corpus against the system's marker library for information just like the ones we present in Section 3.6.2. In this case we use it to evaluate the markers and their performance across the AAAC problems.

As can be seen in the results from the previous section, marker-instances have properties that make them good differentiators for different kinds of document. Here we analyzed 187 documents in at least eight different groupings (essays, novels, poetry, plays, etc.) and find that there are some markers that perform well on the average across all the problems. We present three measures of evaluations: absolute attribution, relative attribution and average Standard Score comparison. Marker #292 has high performance and is listed as the top marker instance by all three evaluations. But the lists are quite different. We believe the relative evaluations are the most relevant for generalizing the effectiveness of individual markers. The average Standard Score evaluation is great for comparison of markers to each other.

We will revisit this problem again in Chapter 4, where we use a weighted multi-marker approach to improve our attribution results and compare and contrast with the individual marker approach.

4.0 CLASSIFICATION WITH MULTIPLE MARKERS

Style-based classification is regular document classification with style markers as features. As seen in Figure 1 (SECTION 1.3), classification has two important components: Distance formula and Machine Learning algorithm. In this chapter, we describe Machine Learning (ML) with a dynamically weighed marker set for style-based classification problems.

Most existing text classification literature (such as authorship attribution) use one or a small set of markers, often subjectively chosen by the researchers for specific problems. We show that for style based attribution, a larger set of objectively derived markers is preferable. We discuss the distance formula and the ML algorithm for our system, as well as experiment with size of the class set in classification problems.

4.1 LEARNING A DISCRIMINATING DISTRIBUTION OF MARKERS (ML ALGORITHM)

There have been efforts to standardize authorship attribution (a notion encapsulated in style-based classification) in the past, most notably JGAAP [Juo09]. JGAAP encoded much of the techniques presented at the AAAC. JGAAP has a 4-prong process for authorship attribution as follows:

1. User loads labeled and unlabeled documents to the system.

2. User chooses “canonizers”, or pre-processing steps to prepare the data. Canonizers include eliminating white space, unifying case and unifying numbers (changing them all to the same token).
3. User chooses “event set”, or features. Examples are “words”, “trigrams”, “parts of speech” or “sentence lengths.”
4. User chooses “distance formula”, for example “RN Cross Entropy”, “Levenshtein distance”, or “Bayesian”.

After these, JGAAP extracts the events and calculates distances between every labeled and unlabeled document and makes an attribution according to similarity. One crucial difference between JGAAP and our approach is that in step 3, only one event set (marker) is chosen to work with, whereas our system works with a distribution of markers. As a result step 2 is also inadequate as different markers may have different canonization requirements. We include canonization routines as parameters to each marker, thus for example, we are able to represent trigrams both with and without counting number symbols. In the following experiment, we show that a multi-maker weighted plurality solution is preferable to a single-marker approach of JGAAP.

4.1.1 Solving AAAC problem A with weighted plurality algorithm

The AAAC contest problem A is one of the most difficult problems out of the 13 problem sets. In fact, the best solution represented in JGAAP attributes nine out of the thirteen documents correctly. This method uses cross-entropy distance operating on the *words* event set with *standard deviation* parameter.

In order to explore the benefits of multiple marker attribution, we experiment with other event sets, keeping the same distance formula. The results are that many other event sets could correctly attribute some number of documents. This number is always less than ten, but documents correctly attributed are not necessarily the same as ones correctly attributed by the *words* event-set. This presents an opportunity to combine the *words* event set with one or more other event sets to achieve a higher score. In such a situation, the event sets would each have to be weighed, approximating the weighted plurality vote algorithm whereby each event votes for a slate of candidates according to its weight.

We conduct 100 experiments with JGAAP version 3 covering all event sets with all parameters with distance formulas "RN Cross Entropy" and "Naïve Bayesian". We keep the canonizing choices constant with only "strip punctuation" and "normalize white space" selected for all experiments. The results are provided in Appendix C. The number of correct attributions for each event-set-parameter-distance-formula combination ranges from zero to nine. (See Appendix C).

In Table 5 below, we list 14 rows from the total 100 sets of trials that we have conducted. The events column in Table 5 contains the shorthand notation for the distance formula ("RN" for "RN Cross Entropy" or "NB" for "Naïve Bayesian"), event sets ("w" for words and "c" for characters), N-gram status ("1g" for unigram, "2g" for bigram, etc.) and finally parameters ("avg" for average, "min" for minimum, "max" for maximum, "std" for standard deviation and "rev" for reverse.).

Table 5. Weighted plurality algorithm votes for AAAC problem A JGAAP events

Documents:	1	2	3	4	5	6	7	8	9	10	11	12	13		
Events														**#Cor.**	**wgt.**
RN-c-1g-avg	1	13	11	8	10	12	9	9	11	10	10	10	9	5	1
RN-c-2g-max	10	7	11	9	11	1	8	4	11	4	11	10	4	3	2
RN-c-2g-min	3	4	3	3	7	3	3	3	3	13	3	3	13	1	1
RN-c-2g-rev	1	4	11	1	1	1	1	1	1	4	1	1	1	3	1
RN-c-3g-std	3	13	3	7	3	3	13	3	3	13	3	13	7	3	1
RN-c-3g-min	3	1	3	7	3	3	13	3	3	13	3	13	7	2	1
RN-c-4g-std	7	13	3	7	13	3	13	7	3	13	13	13	7	2	1
RN-w-1g-std	12	2	10	4	10	12	8	1	5	4	6	2	9	9	3
RN-w-1g-max	6	1	4	4	4	2	8	4	5	4	6	9	9	5	1
RN-w-2g-std	9	13	10	10	10	12	10	1	5	4	10	2	4	7	1
RN-w-2g-max	9	1	11	10	9	4	10	1	5	5	1	4	9	4	1
RN-w-4g-std	9	13	4	7	5	12	5	2	9	10	6	11	9	5	1
NB-w-1g-std	6	7	6	11	7	10	13	6	7	6	6	8	2	1	1
NB-w-3g-std	3	10	9	7	10	12	7	10	9	5	5	11	10	4	1
Correct	3	13	11	7	10	12	8	1	5	4	6	2	9	13	

In the first row of the table, we can see the documents numbered one through thirteen. The last row is the correct authors associated with each document. The performance of some of the better event set / parameter combinations are listed in each row. Each green cell also signifies a correct attribution. The "# Cor." column sums up the number of correct attribution. The event "RN-w-1g-std" has the best performance with nine correct attributions that is the top individual performer for problem A. It misclassifies the first four documents. However there are some event sets that can classify some of the first four documents correctly, while misclassifying some of the later nine.

We find that with a simple weighted plurality algorithm we can classify thirteen out of thirteen documents correctly. The weighted plurality algorithm is designed to construct a compound formula from a finite set of weighted predictions. It gives a specified number of "votes" (weight) to each predictor and then counts to see which prediction has a plurality and that becomes the overall prediction of the set.

In this case, each of the event sets is considered a distinct prediction algorithm. The weight distribution for each event row is marked in the last column, titled "wgt." As we can see, almost all the weights are 1, except for two rows. Counting all the votes now in each column allows for all thirteen documents to be classified correctly using these 14 event sets.

While this simple experiment is a post-facto analysis of the event set performance in one problem of the AAAC, it demonstrates that the power of multiple predictors working together via the weighted plurality algorithm.

4.1.2 Solving the Marker weight optimization problem using modified first-choice Hill Climbing

Our system uses a modified first-choice Hill Climb algorithm to optimize a weight vector for all the markers involved. This is a slightly different approach to the one demonstrated above with the AAAC problem A and it is geared toward a larger set of markers (hundreds). It is a similarity-based method suitable for profile-based attributions [Sta09]. As our focus is not on ML algorithms, this one is being presented here as the default choice we made for our system and not a statement of it being the

best possible algorithm for this problem. Our system is flexible enough to take any ML algorithm and in fact, this one could be replaced by another one of the user's choice. We demonstrate this algorithm's performance using the Brown corpus. The algorithm is described in pseudo code below.

```
Define bestScore (M,Wi,D):

      Wn = Wi
      Score = 0

      for I in 1...ITERATION_MAX:
          m = random(M)
          Wr= randomAdjust(Wn,m)

          if (Attribute(D,M,Wr) > Score):
              Score = Attribute(D,M,Wr)
              Wn = Wr

      return Score
```

Figure 5. Pseudo code for the modified first-choice Hill Climbing algorithm

Variables used in Figure 5 are described below:

- *M* is a vector of markers in normalized Z score,
- *D* is a vector of labeled and unlabeled documents,
- *Wi* is the initial marker weights, a distribution over markers M,
- *Wn* is the current best weight vector,
- *Wr* is the vector adjusted due to the increase/decrease of a random marker, m,
- *Score* is the number of correct attributions, calculated by Attribute() using the Euclidean distance formula.

Our system used the algorithm above to learn the weight matrix on the Brown corpus, which is divided into 11 genres. The performance on a randomly selected test set of 33 documents (3 from each genre), can be seen in Figure 6. 12 experiments (4 sets of markers and 3 trials per set) were conducted each with 2 fold cross validation. Multiple sets of makers were used, each cumulatively larger in number than the ones before, to also demonstrate the positive correlation between increasing accuracy and larger sets of markers.

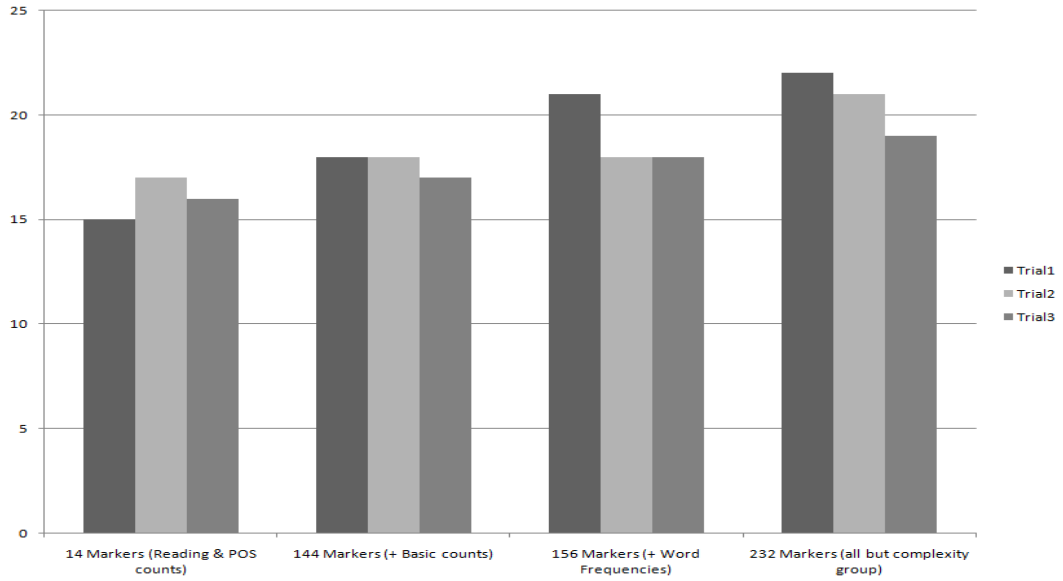

Figure 6. Brown corpus classification using modified first-choice Hill Climbing: The Y-axis is the number of documents correctly attributed, which, as shown, is correlated positively to the number of markers used.

4.2 STYLE VECTOR DISTANCE FORMULA

$$Classify_C(doc) = ARGMIN_C(D_{M,W}(doc, C))$$

$$D_{M,W}(\alpha, \beta) = \sum_M W_M \sqrt{(M_\alpha - M_\beta)^2}$$

Although any user-defined distance formula can be used, our system and most of our experiments use the Euclidean distance formula with shortest-distance indicating the most similarity. The distance and attribution formulas are as follows:

Classify() is the function that associates a document, *doc*, with one of the corpora from the corpus set *C*. It works by finding the corpus with the lowest distance between *C* and the *doc*. The second line describes the distance formula itself. Two input entities (either a document or a corpus), along with a marker matrix, *M*, and a trained weight vector, *W*, are passed in as parameters. The distance between any marker in one corpus and its counterpart in the other corpus are found by the Euclidian distance (square root of the difference, squared). The overall distance multiplies each individual maker distance by the marker's weight, and sums up across all markers.

4.3 INCREASING THE NUMBER OF CLASSES

In this section we examine the effects that the size of the target class set (number of distinct possible styles) has on attribution performance and markers. In particular, we explore the persistence of some markers across a set of experiments of progressively increasing number of styles.

4.3.1 Effects on training convergence as classes increase

We use the AAAC corpus problem A which has thirteen distinct classes, each representing a different author. The original training set provided for the contest has three short essays per author. Thus to distinguish between all authors, we need the training to produce a weight vector that can correctly classify as many training documents as possible (convergence).

We begin the series with an experiment that only includes two classes: Authors 1 and 2 and we converge on six training documents being classified correctly. The experiment is repeated *n* times, in order to get variations of the weight matrix and an average convergence in terms of the percentage of training documents correctly classified.

After the n runs, we increase the number of classes to three, and consequently the number of training documents by three and run the next experiment n times as well. We continue progressively adding a class, until all 13 original authors have been included. The convergence results are shown below.

Figure 7. Training convergence averages with increasing number of classes

The experiment was run n = 4 times per class-set, resulting in a total of 48 runs. Each run was a fixed number of markers (m = 530), and fixed number of cycles (c = 106,000) of the modified First-Choice Hill-Climb algorithm.

Although we can generally see the convergence drop after the 4-class experiments, it is uneven, probably due to two factors: The first is quantization. The relative percentage contribution of a single document diminishes as the classes are increased. For example, the difference between the 4-class and the 5-class experiments in convergence is a single document. The 5-class experiments, on average, classified 14 out of 15 documents correctly.

The second factor is the relative difficulty of the additional class that is added at each step. The new training documents to be added to the mix, are not all of the same quality and consequently have different effects on classification performance.

4.3.2 Effects on discriminating style markers as classes increase

In this section, we examine the effects of increasing the number of classes on markers. Specifically, we are interested in seeing how many of the important markers remain the same across the experiments. The same 48 experiments from 4.3.1 are examined for marker consistency.

Each experiment produces a new marker weight distribution, *W*. Recalling the actual algorithm, each cycle allows for one or more markers to be elevated in weight and the resulting vector is examined for classification of all the training data. Thus, at the end of each experiment, the final *W* vector can be divided into two groups of markers: Those whose change in weight leads to a new high score, and those who did not. The members of the first group have varying weights assigned to them, which typically differ in each run, even with the same exact parameters and classes. However, the weights of the members of the first group are always higher than the second group. The rest of the markers in *W*, which are in the second group, have all the same low (lowest) weight in the *W* vector. Therefore, we can conclude that any marker in the first category, which by definition has made a difference in attribution score, has a direct effect on performance and is more important to the classification effort, regardless of relative weight (which is at least guaranteed to be higher than the lowest class).

With this information in mind, we map all 530 markers and divide them into the aforementioned two categories for each experiment: Contributive or non-contributive. Within the n trials of each experiment, we keep track of all (superset) contributive markers, and after the trials we maintain a list of markers that have been in the contributive category at least once during the trial. We then examine the longer-term trend as we increase the number of classes in the experiments. In the following graph, a bar is shown for every marker that has been a contributor in at least one experiment. Multiple contributions in the same class-set are not counted multiple times, meaning a maximum of one point is given to a marker that has made any appearance in any of the n runs of the contributive group.

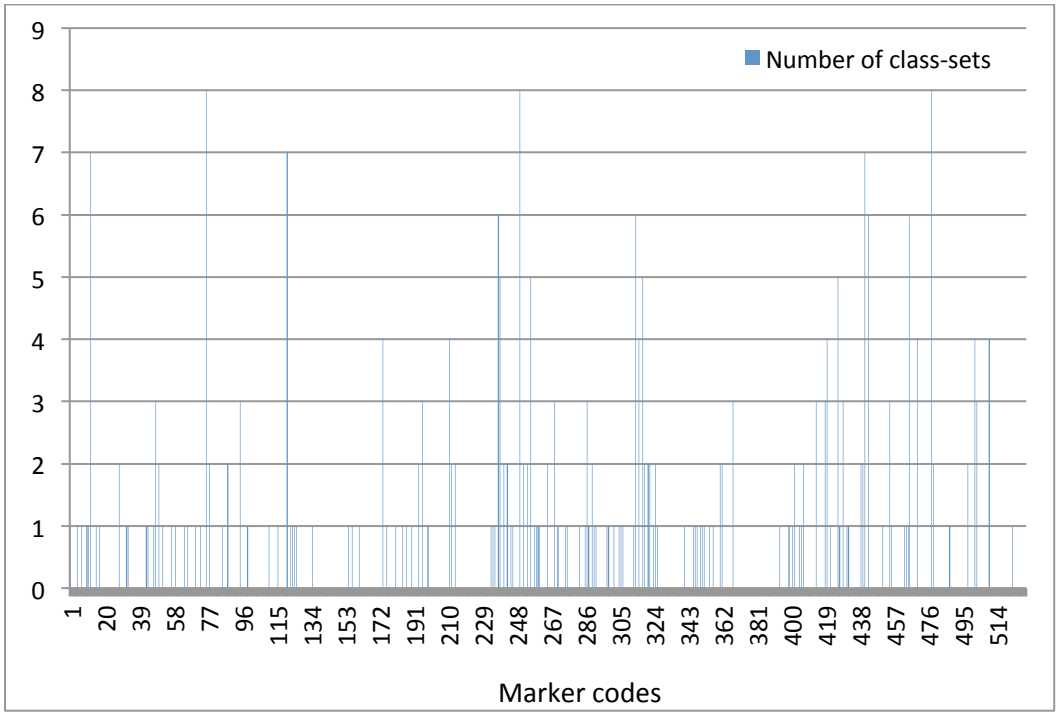

Figure 8. Marker contributions across classification experiments of different sizes

The first notable feature of the graph is the large number of markers that are not contributive to any experiment. Not even one out of the 48 x 106,000 cycles saw their contribution to *W* result in one additional correct attribution.

The next largest group is those markers that have made exactly one contribution to an experiment. These markers, which have made a scoring difference in only one class-set size experiment, could not be considered part of a trend or examined for consistency. We break down the entire distribution of the 530 markers below.

Table 6. Distribution of contributive markers

Number of experiments with distinct class-sets	**Number of contributive markers**
0,(no experiment showed these markers as contributive)	385
1	88
2	25
3	11
4	7
5	4
6	4
7	3
8	3

Discounting the majority of the 530 markers (385 which never contributed and 88 which contributed only once), we are left with a group of 57 markers that can be considered "universal" to some degree, as a consequence of their persistence in contribution in experiments of increasing class-sets. The most common markers made contributions to 7 or 8 distinct experiments. We describe those markers more specifically below.

Table 7. Most persistent markers in the AAAC problem A series of classification experiments

Marker # (class-sets)	Marker	Method of derivation
#11 (7)	frequency of all other POS tokens	number of all part-of-speech tagged words other than nouns, verbs, adjectives, adverbs, pronouns, and determiners, divided by total number of words in a corpus
#75 (8)	complex words per document – mean	the average number of words of 3 syllables or more in each document
#120 (7)	Diction variations per document	the total number of distinct texts that can be created from all GNU Diction rule-set substitutions that are possible per document
#249 (8)	punctuations per sentence – minimum	the lowest number of punctuation marks found in a sentence in a document
#440 (7)	top100words-28	frequency of the 28th most frequent non-common word in English
#477 (8)	top100words-65	frequency of the 65th most frequent non-common word in English

5.0 TRANSFORMATION

Style *transformation* refers to the process of changing a document associated with one style into one associated with a different style, while preserving the overall meaning. This process can have a target (desired style), or an anti-target, meaning the goal is to *obfuscate* and move away from the initial style, but not necessarily toward another particular one. Transformation systematically applies a series of *transforms*, each consisting of a way to rewrite part of the text at sentence or paragraph level. Each transform application is adopted if the resulting document is measured to be statistically closer to the target style.

5.1 OVERVIEW

The transformation process uses a set of independent, modular and heterogeneous monolingual text-to-text translation routines (transforms) working together to produce alternative versions of the text that preserve correctness and semantics of the original text. One or more of these alternative versions of the text are then used in a style classification engine that will determine if the new version of the text meets the style goals of the overall system. If the new replacement moves the document closer to its target style, then that replacement is adopted as the latest working copy of the text.

Section 5.2 covers in detail the individual transform algorithms currently used in our system along with experiments and discussions.

5.1.1 TARGET AND ANTI-TARGET

As seen in Figure 2, one of the input corpora used for classification can be designated as a "target", meaning that the process aims to move the document closer and closer to the style exhibited by that corpus. This, however, is not a necessary designation. We discussed the "style obfuscation" use case in chapter one. In this use case, the target is not designated and the primary goal is to be no longer associated with the original corpus. The original style in this situation is called the "anti-target".

5.1.2 TYPES OF TRANSFORMS

For this project and our conception of transforms as independent, modular, and heterogeneous, transforms can be classified algorithmically and computationally in many different ways. For example, some may require full linguistic parsing, others chunking and still others may have very little to do with linguistics or Natural Language Processing. All transforms produce responses that are scored regardless of how they were produced or which algorithm was employed. Thus the only point of distinction we find useful is the scope of the text a transform operates on, namely: document level, paragraph level and sentence level transforms.

Sentence level transforms are the most common kind that we employed in this project. They consist of a single sentence (not necessarily defined in the linguistic sense) as input and they produce one or more alternatives to the same sentence as output. The

goal for the transform is to produce alternatives that preserver correctness and semantics of the original. In order to accomplish the goal, transforms can access other databases and tools that we will elaborate on later. A detailed description of each sentence level transform used in this project is found below in Section 5.2.

Paragraph and Document level transforms are similar to sentence level transforms except that they accept a paragraph and a document as inputs respectively. For this project we did not implement any document level transforms and only one paragraph level transform. The paragraph level transform is called "paragraph adjust" and its function is to either split or combine paragraphs in order to mimic the per document paragraphing style of the target corpus.

5.2 STYLE TRANSFORMS

Style transformation is the systematic application of a series of style transforms. Each transform is a simple monolingual text-to-text generation routine with two simple goals: 1) Preservation of the level of correctness and grammaticality of the original text and 2) Preservation of the original semantics as much as possible.

To demonstrate the viability of style transformation, we contribute several transforms as well as adopt others in the field. We discuss these in detail for the rest of this chapter. Each subsection below is dedicated to a single transform. The first part of each subsection contains a quick-reference table that outlines some basic facts about the transform and how it operates.

5.2.1 ACTIVE-TO-PASSIVE VOICE STYLE TRANSFORM

Table 8. Active-To-Passive transform

Transform	Active-to-Passive (A2P)
Type	sentence level
Theory	active voice sentences with transitive verbs and objects can be directly transformed to passive voice equivalents
Data sources	initial 20 basic rules from EnglishPage.com then mutated to about 120 rules, adding various possibilities for particles, adverbs and complex noun phrases
Technologies	Link Grammar Parser, Nodebox, NLTK
Example	*The boy kicked the ball. => The ball was kicked by the boy.*
Development	Foaad Khosmood, implemented as Python class

The active to passive voice transform performs two operations. First, it tests a sentence for being active. This is accomplished by running through a library of rules that denote the various formats of active sentences. Second, if the sentence is deemed to be active, it reformulates it in the passive form. In this section we briefly cover the linguistic background and describe the transform in detail.

5.2.1.1 Active and passive voice

The classic formulation for a simple active voice sentence [Lut05] is:

[subject] + [verb] + [object]

For example "Johnny hits the ball."

[Johnny] + [hits] + [the ball]

In the passive voice, the main occurrence is that the subject phrase or actor is de-emphasized and becomes passive. The object phrase becomes the subject (is acted upon) of the new sentence [Eng10]. The verb is preceded by the tense-preserving version of "to be" and transformed into its past participle form.

[subject being acted upon] + [be] + [verb: past participle] + [by] + [actor]

Applying the rule to our original example:

[The ball] + [was] + [hit] + [by] + [Johnny]

It should be noted that most sentences are active [Fow26]. Furthermore, large portions of the typical passive sentence constructions do not contain an agent [Lut05]. For example "The ball was hit" or "Lunch is served". In order to transform an active sentence into passive, it must contain an agent, which is very common. However, when transforming sentences into their passive form, we can encounter correct but awkward phraseology [Fow26] that is generally less commonly used.

5.2.1.2 Processing active sentences

When attempting to process the original active version, we can identify the three constituent phrases with three variables of 1, 2, and 3.

Table 9. Active and passive voice constituents

Constituent	Example	Variable #
actor	Johnny	1
verb	hits	2
object	the ball	3

Substituting variables, as well as using the Link Grammar (2003) parsed notation (Penn Treebank) we rewrite the active sentence pattern as:

```
(S (NP 1)
    (VP 2
        (NP 3)))
```

And the corresponding passive sentence would be:

```
(S (NP 3)
    (VP [to be] 2
        (PP by
            (NP 1))))
```

[Eng10] identifies 20 different verb tenses and sentence constructions that augment this classic pattern. We encode these rules and add a few of our own for the initial transformation set. A few examples of the initial set of active patterns for the verb phrase (variable 2) are listed below. These rules will be augmented to cover many more cases.

Table 10. Example of initial active detection rule patterns

Tense	**Pattern**	**Example**
Simple	(VP 2 (NP 3))	I hit the ball.
Present	(VP am|is|are (VP 2 (NP 3)))	I am hitting the ball.
Past continuous	(VP was|were (VP 2 (NP 3)))	I was hitting the ball.
Simple future	(VP will (VP 2 (NP 3)))	I will hit the ball.
Future perfect	(VP will (VP have (VP 2 (NP 3))))	I will have hit the ball.

5.2.1.3 Functionalizing necessary information

To do a full active to passive transformation, all that remains is to determine the morphology of the individual variables. For example, if the actor is a personal pronoun such as “he”, it would be converted to the object version of the same which is “him”. “I”, ”we”, “she”, ”they” would be replaced with “me”, “us”, “her”, “them” respectively. The following table outlines the functions that are needed to operate on each variable to determine the final correct passive voice form of the variable.

Table 11. Functions needed for grammatical passive construction

Variable	Functions needed	Purpose
1, actor	object form is needed	Final form of NP
2, verb	past participle form?	Correct form
3, object	plural?	Determines the form of [to be]

5.2.1.4 Extending the pattern set

We begin with the initial 20 or so rules discussed in Section 2. These rules make several assumptions for the sake of simplicity that result in a lower match rate when examining the text for them. These assumptions include having simple verbs without particles and without modifying adverbs, and simple noun phrases as subjects.

In order to have a more general grammar, covering some of the missed cases, we mutate the existing rules by adding features to them, and adding the new version to the rule set. Specifically, three features are added using this mutation technique:

1. Support for verbs with particles such as “gave you up” or “drive over it”. Multiple variations of the particle are automatically added to each existing rule form.
2. Support for adverb phrases, such as “slowly driven”, or “shouting vehemently”. Adverb phrases can appear before, after or after the particle.
3. Support for more complex subject noun phrases such as “attorneys general” or “crime of the century”.

Adding these mutations, about 9 in all, would extent the rule set to about 180 rules from the original 20.

5.2.1.5 Transform design

The main loop of the transform is shown in the following figure.

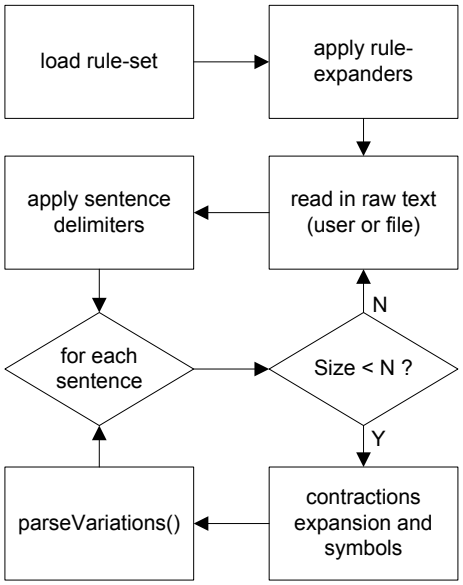

Figure 9. A2P high-level operation.

The rule set is loaded and expanded. The reader module reads a sentence in raw text form passed into it as a parameter. Several small pre-processing steps are completed, including regular-expression based contraction expansions and policy enforcement on the size of any sentence (set to 20 words by default). The major work is done inside the ParseVariations() routine that we discuss below.

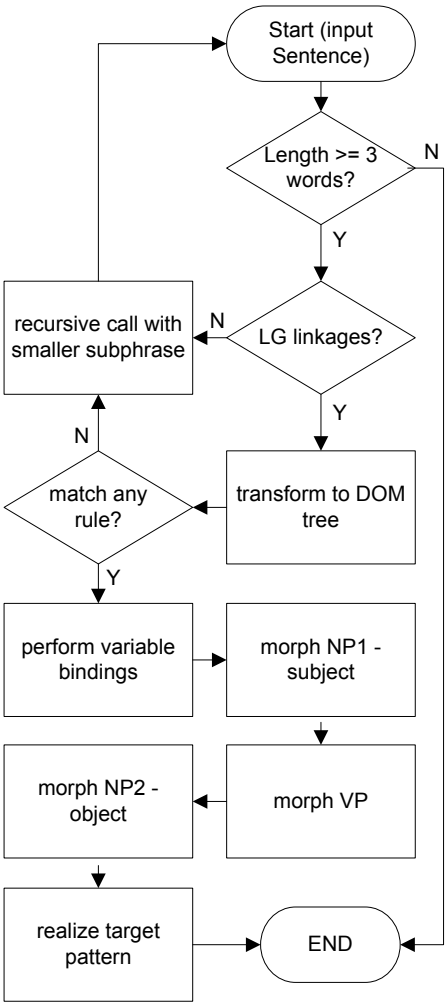

Figure 10. ParseVariations()

ParseVariations is the main routine operating on a single sentence and either producing a valid passive version, or returning "False" otherwise. The routine is called so because it recursively explores sub-phrases of the given input sentence, looking for matches to the active pattern set. ParseVariations() begins by checking the size of the input sentence. Any sentence below the size of 3 words is automatically discarded and

"False" is returned. If the sentence is of the appropriate size, then it is passed to Link Grammar [Sle93], for parsing. If Link Grammar cannot find any linkages, and thus cannot produce a constituent tree, we consider this also a failure. It should be noted that Link Grammar can reject well-formed sentences and thus at present this is a limitation for greater processing of sentences.

5.2.1.6 Applying the rule set

Assuming at least one linkage is found, the resulting Link Grammar constituency tree is turned into an XML Document Object Model (DOM) tree. This step is necessary in order to compare the sentence with the library of active voice patterns. Each entry in the library consist of a list of 4 items: a "pattern" to be matched by the active voice sentence, a second pattern indicating the passive version of the same sentence with same variables placed in the appropriate passive position, a function name to be called to handle the morphology of the variables and lastly parameters to be passed into the function. The listing below shows the simplest rule in the library as a Python list entry. This one handles the "simple present" transformation case.

```
[ "[S  [NP 1 NP]  [VP 2 [NP 3 NP] VP]  S]",            #active "pattern" to match
  "[S  [NP 3 NP]  [VP 2 [PP by [NP 1 NP] PP] VP]  S]", #passive pattern to gen.
              active2Passive,                          #name of function to call
              'SimpPre'                                #parameter (simple
present)],
```

Figure 11. Python list entry for one A2P rule.

As can be seen, the first entry is an active matching pattern in quotes, in slightly different form than we examined in Section 5.2.1.2, but it is in fact the same exact active/passive patterns expressed there. The third entry is the function call and the

parameters (beside the sentence details and bindings) that are passed with it. In this case the flag 'SimpPre' helps the active2Passive() routine achieve the morphological requirement on the variables.

Each entry in the rule library is examined for matching purposes. In order to perform a structural isomorphism check, the pattern will also need to be converted to the XML DOM, as the sentence itself was. Once both sentence (Sdom) and pattern (Pdom) are in DOM form with the familiar Link Grammar "S" (sentence) node as their respective roots, then we call a recursive routine, that systematically checks to see if the two structures are equal. Similar to tree-isomorphism, the routine first checks to see if the current root nodes are equal (or one is an unsigned variable). If they are equal, then the number of children of each root is compared. If those are equal, then the children themselves are compared one at a time recursively until there is either a mismatch or everything has been compared and therefore the two structures are equal.

5.2.1.7 Modifying the noun phrase variables

At this point, a rule has been found to match the sentence structurally. The next step is to bind the variables that are present in the active pattern portion of that rule. During the binding process, the sentence DOM is traversed one more time recursively and a grown Python dictionary of variable/bindings is returned.

The function specified within the rule, "active2passive" is now called with the pattern, sentence, flags and the bindings dictionary. The task is to realize every variable

in the action part of the rule in the correct morphology. Most rules have only three variables: N1, V and N2.

N1 is the original active sentence's first noun phrase, the subject, or the "actor" [Eng10]. Processing this variable is relatively easy. Either N1 is a personal subjective pronoun like "I" or "they", or it contains other pronoun like "our cat", or it is an object like "the tank" or "my dog". Only in the first case is any morphology necessary, and that task is easily accomplished by looking up the objective version of the pronoun. We have specified a useful lookup table for this and other situations. The table is reproduced below as a Python list.

Table 12. English pronoun forms. "subj." = "subjective", "obj." = "objective" and "pos." = "possessive".

person	plural?	subj.	subj.	obj.	obj.
1	0	I	my	Me	mine
2	0	you	your	You	yours
3	0	he	his	Him	his
3	0	she	her	Her	hers
3	0	it	its	It	its
1	1	we	our	Us	ours
2	1	you	your	You	yours
3	1	they	their	Them	theirs

In the vast majority of cases with NP1, we simply look up to see if NP1 is in any column 3 of the pronoun table. If yes, we replace it with the entry in column 5 of the same row.

NP2 is the object in the original active construction. It typically leads the passive sentence. If it's a pronoun, the reverse process of NP1 must take place, i.e. a lookup in

Table 12 column 5 and if successful, replacement with entry in column 3. For example: "The dog is chasing *them*." => "*They* are being chased by the dog."

The biggest challenge with NP2 is named entity recognition. Quite often Link Grammar does not recognize a multi word phrase as a noun phrase, and thus tries to parse the structure further, creating mismatches with the existing rules. We do a number of tricks to try to mitigate this situation. We check WordNet (2008) and another external named entity library. The other important point on NP2, is its singular/plural status. This information is necessary in order to use the correct form of the "to be" verb in the passive sentence. We again use WordNet and some external libraries to determine plurality, including one source for irregular plurals such as "attorneys general" or "parts of speech". In addition, we specifically check for grouping words such as "a pair of trousers" (singular) or "a couple of sheep" (plural). If none of the libraries or special cases applies, we collapse down to lexical methods for plurality determination.

5.2.1.8 Modifying the verb phrase variable

Lastly, the verb phrase (VP) needs to be replaced with the past participle version of itself, and be augmented by other features depending on the tense that was matched by the rule.

Most verbs' past participle is the same as their simple past tenses [Lut05], for example "run"/"ran" and "shave"/"shaved". Some irregular verbs have past participles that are different from their simple past tenses, for example "eat"/"eaten" or

"see"/"seen". Much like the previous cases, the irregular morphologies are checked for first and if there was no entry, a simple past version is used.

The "to be" verb in the passive sentence is modified depending on the plurality and pronoun version of NP2, for example "He *was* taken by his father," or "The pigeons *are* fed by me" or "I *am* hated by the law."

If the VP is negated, which we specifically check for, then the word "not" is inserted after the "to be" verb. Additional insertions of words such as "have", "going to" and "being" are added strictly according to the specific tense rules, after the "to be"/"not" terms but before the actual VP in past participle. Lastly, the rules augmented with verb particle support, include an additional variable "PRT" which is inserted immediately after the VP without modification. Examples are "I *dressed* you *up*." => "You were *dressed up* by me," or "He *builds up* his business." => "His business was *built up* by him."

5.2.1.9 A2P evaluation using the Tatoeba corpus

Tatoeba [Tat10] is a giant user contributed sentence level translation project containing millions of simple sentences in dozens of languages. The sentences are peer reviewed and tend to be of high quality of accuracy. The project allows one to download all sentences of any language. At the time of our access the English language sentences numbered more than 150,000. We decided to use subsets of the Tatoeba English database for our evaluation.

It is difficult to evaluate very large sentence reconstructions without possessing large human resources. Lacking this, for evaluation of the actual A2P functionality, we relied on a 2-prong strategy. First, we run Link Grammar on the resulting passive constructions to see if they produce linkages. If such linkages are produced, it is highly indicative of correct construction of the sentence. However, as we quickly realized, Link Grammar rejects many good constructions as well. So, the second part of our strategy is to human-verify a selection (if possible the entire set) of the negative sentences as indicated by Link Grammar. Using machine translation approaches, we basically tag each sentence with "correct", "incorrect" or "acceptable". Examples of each class are available below.

5.2.1.10 Evaluation experiments

We are interested in how many sentences are examined and out of those, how many are matched by our rules, and finally what was the correctness rate of the ones we matched and transformed. Since we have a policy parameter restricting the size of the eligible sentences, we decided to use 4 different parameters to compare and contrast accuracy results. These are 10, 8, 6, and 4 word-length-maximum sentences respectively. The other parameters (sub-phrasing, and use of extended rules) were both active for this experiment.

Table 13. A2P experiments

Max Words/Sentence.	Total Examined	Total Active matched	Total Passive parsed	False Negatives

10-words	3457	514	313	30
8-words	4079	476	295	30
6-words	7953	491	367	25
4-words	16350	216	179	19

We calculate two ratios, one attempted ratio, and other accuracy ratio. The accuracy ratio takes into account the false negatives and adds that number to the "total parsed" number by Link Grammar before dividing it into the "total Active matched." We observe that accuracy rises dramatically with the smaller sentences, starting at 66% with 10-word-maximum sentences and passing 91% on 4-word sentences. Naturally, selectivity goes down with smaller sentences, given that much less of the corpus is now eligible to be considered for the transformation.

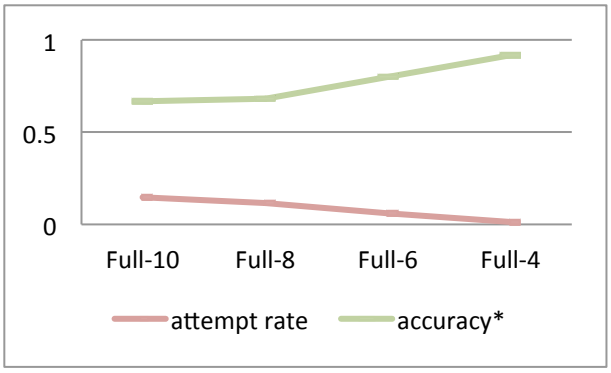

Figure 12. Rising accuracy with smaller sentences. *Accuracy is out of 1.0 = 100%

5.2.1.11 Example transformations

We consider individual sentence results in three categories: correct, acceptable and incorrect. We generally observe that passive sentences able to be parsed by Link

Grammar parser are correct. Of the ones not parse-able, however, some can be categorized as acceptable and some even correct. We briefly examine sentences in each of these categories. Starting with the correct category, the resulting sentences were successfully parsed by Link Grammar:

1. My little brother is watching TV. => TV is being watched by my little brother.
2. I like my job very much. => My job is liked very much by me.
3. Sometimes he has difficulties with being articulate about his views. => Sometimes difficulties are had by him with being articulate about his views.
4. The orchestra makes discordant noises when tuning up. => Discordant noises are made by the Orchestra when tuning up.
5. He told me the story of his life. => I was told by him the story of his life.

By contrast these next few sentences are failed cases:

1. I was rereading the letters you sent to me. => The letters were being reread by me you sent to me.
2. They are making too much noise, I can't concentrate. => Too much is being made by them noise, I can not concentrate.
3. I knew it was plastic but it tasted like wood. => It was known by me was plastic but it tasted like wood.
4. He will make you eat dirt. => He will make dirt is eaten by you.

And we tagged these as acceptable, even though Link Grammar rejected them. In most cases, there is no other passive construction possible for them:

1. I have a dream. => A dream is had by me.
2. But you never told me this! => But I was never told by you this!
3. It costs and arm and a leg. => An arm and a leg is cost by it.
4. Everyone has strengths and weaknesses. => Strengths and weaknesses are had by everyone.
5. I wish you had told me the truth. => I wish I had been told by you the truth.

5.2.2 DICTION TRANSFORM

Table 14. Diction transform

Transform	Diction
Type	sentence level
Theory	suggestions from various "good style" writing tools such as GNU Diction can be implemented as transformation rules
Data sources	1. 655 rules from GNU *Diction* v. 1.11 2. 658 "wordiness" rules from Steve Hanov [CITE] (Stevehanov.ca) partially based on William Zinsser's *On Writing Well* 3. 409 word/alternative pairs from *The A to Z of alternative* words by the Plain English Campaign [CITE]
Technologies	Python
Example	*The fact is, in the majority of cases, X is defined as an unknown. => Usually X is an unknown.*
Development	Foaad Khosmood. Implemented as a Python class *DictionOps*.

The Diction transform takes advantage of a number of open-source rule sets, or rules of thumb, that exist in the writing composition and education communities. These rules are originally written for the benefit of writing assistance programs and are meant as mere suggestions as opposed to bona fide exact replacements. Therefore, under some circumstances, their implementation results in ungrammatical sentences. We have created string-to-string replacement rules from the suggestions provided in the sources.

We present *DictionOps* as an example of a transform easily implementing rules from existing text-to-text dictionaries such as the three that we currently use.

5.2.2.1 Rule sources

Roughly 1700 rules (counting possible overlaps) have been implemented from three sources already introduced above.

5.2.2.1.1 GNU Diction

Diction [Haa97] is a writing assistance program that prints out all finds "sentences in a document that contain phrases from a database of frequently misused, bad or wordy". Roughly based on The Elements of Style [Str18], it has a rule library with suggestions that come in four flavors:

1. Straight string replacements such as "a lot of" => "many"
2. General advice like "cliché" or "avoid"
3. Suggestion to rewrite the sentence or start a new sentence
4. Reference to another word such as "see xxxx"

Of these, only the first and the fourth are usable by our system. Our rule dictionary does store all the information from the original rule set but at the moment only straight replacements are implemented. References to other words are resolved after the first pass in reading the rules and if they lead to a replacement, they are adopted as another rule.

5.2.2.1.2 Steve Hanov

Steve Hanov has created a set of words and expressions to generally reduce "wordiness" in essays. He has based on work on the William Zinsser's *On Writing Well.* The rules come in only two flavors: replace with nothing or replace with the suggested

words. While a few still depend on further human examination, the majority of the rules are very machine friendly, and adoptable as straight string replacements. We add these rules to our rule dictionary after loading the *Diction* ones.

5.2.2.1.3 Plain English Campaign

These rules are from an online publication, *The A to Z of alternative words* [PEC01]. They are presented as an alphabetically sorted set of patterns (usually words, but sometimes short expressions) with alternative suggestions for each. A fair number of the suggestions are ambiguous or otherwise not directly replaceable. We adopted a subset of these rules chosen by examination and added them to our rule dictionary.

5.2.2.2 Transform design

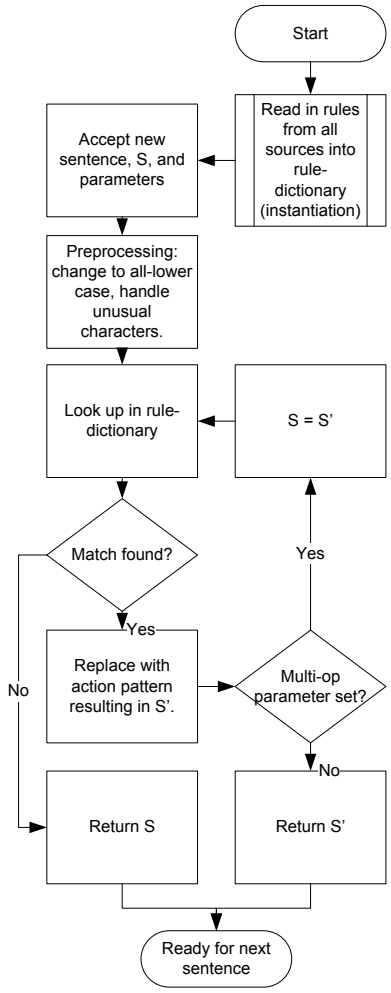

Figure 13. Diction transform design

During instantiation of the class, the constructor reads in the rules from all sources. These rules are deliberately left in general text files to allow easy updating and modification by humans if necessary. As the rules are read in, the ones for duplicate patterns are added to the one pattern's alternative list. This way, there are unique keys to access the dictionary with, but the additional suggestions are not discarded.

The class object then waits until the "xform()" function is called. At that time, the function accepts a sentence and some parameters and proceeds to systematically search all its rules for a possible match. The accessing of the rules is prioritized according to the size of the pattern, that is, the longer patterns are examined first. As there may be multiple rules applicable to a single input sentence, a parameter called "multi-op" is passed in at call time. If set to true, this parameter allows the rule search to continue even after a previous rule has been found and applied to the sentence.

Lastly, if multiple suggestions are available for a matched pattern, multiple replacements are made and multiple sentences are returned to the calling object.

5.2.3 NODEBOX TRANSFORM

Table 15. Nodebox transform

Transform	Nodebox
Type	sentence level
Theory	fixing misspellings or alternative ways of expressing numerical phrases can reflect stylistic choices
Data sources	WordNet, small table of numerical words
Technologies	Python, Nodebox Linguistic suite, NLTK, Stanford part-of-speech tagger
Example	*We recieved 6533 boxes. => We received thousands of boxes.*
Development	NodeBox Linguistics module by Frederik De Bleser and Tom De Smedt. Foaad Khosmood implemented front end, explicit numerics and WordNet checking.

This transform uses two important functions spelling and quantification, from the NodeBox research project's "Linguistics" module. The transform relies on these

functions to make sentence alterations based on spelling and quantification of numerical expressions.

5.2.3.1 Spelling

NodeBox offers two spelling functions, one called "suggestion" where NodeBox is ambiguous about the possible misspelling, and the other called "correct" where NodeBox is more certain that an unambiguous replacement can be made. The "correct" function, however, applies very infrequently and still can make inaccurate corrections. In order to increase the accuracy and applicability, our module does a separate check for words in WordNet, with the idea being that a word that has an entry in WordNet is probably spelled correctly already. Words that are part-of-speech tagged as proper nouns or named entities are not checked. If the word has an English suffix or prefix, those are stripped off first. This is because not all suffix and prefixed words are separately represented in WordNet. If the word or it's stripped version is still not in WordNet then the word is passed to the NodeBox correct() function. The result of this function may be a corrected version or the same as input.

5.2.3.2 Quantification

The Nodebox quantify() function accepts a singular noun such as "goose" and an integer such as 2 as inputs. The output is a quantification expression such as "a pair of geese." NodeBox has a fairly extensive list of quantification terms and irregular plural nouns that it draws on for this function.

Our module checks all numeric/digital words that are followed by plural non-proper nouns. We derive the integer from the numeric word and use NodeBox's own singular() function to turn the noun into its singular equivalent. We then call quantify() and replace the two words with the results.

5.2.3.3 Transform design

The input sentence is tokenized and part-of-speech tagged. Each of the functions spelling() and numerics() are called in turn and the results are returned to the calling object. The implementation of both of these functions is explained in the figure below.

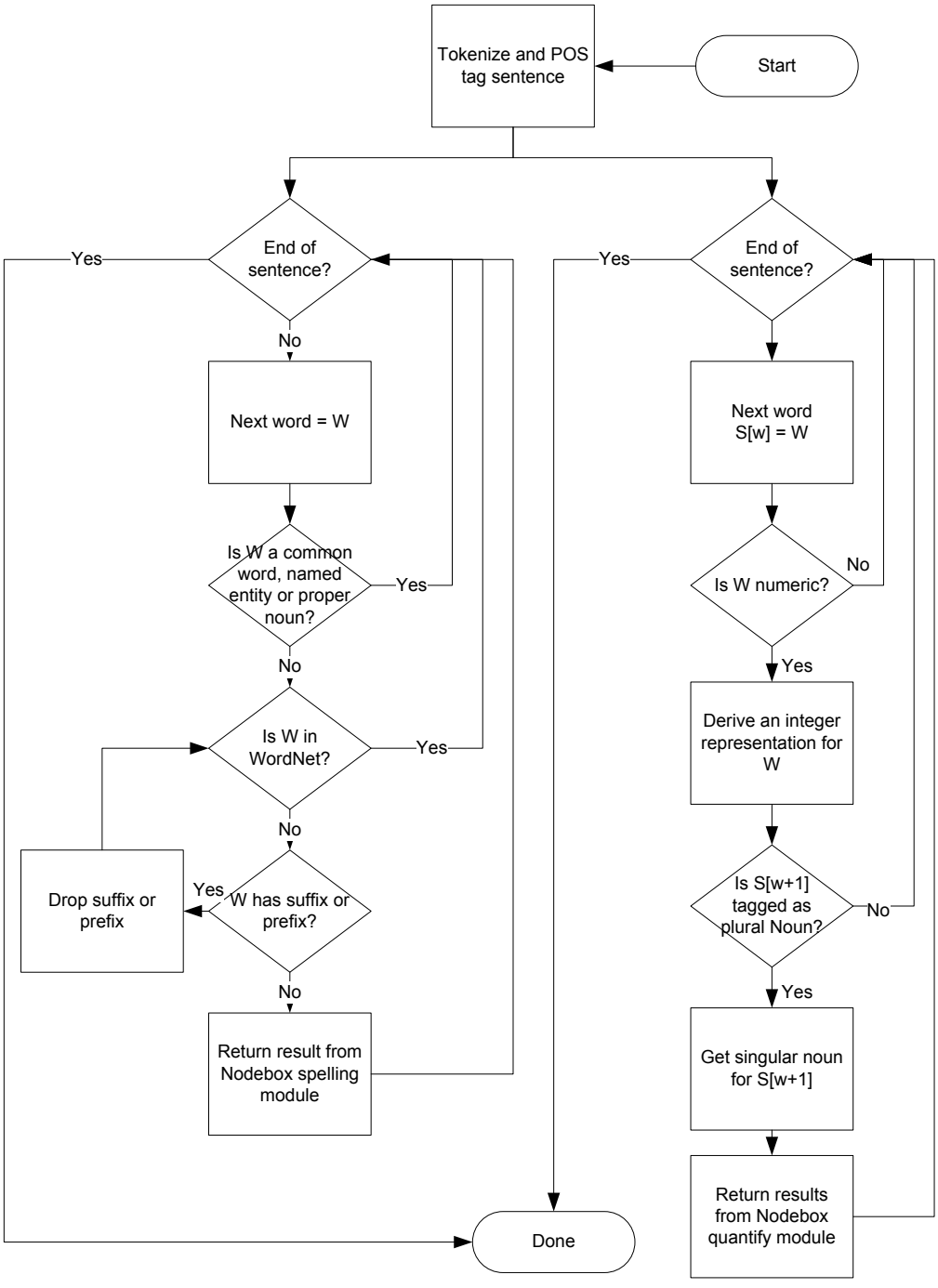

Figure 14. Nodebox transform design

5.2.4 PAROPS TRANSFORM

Table 16. ParOps transform

Transform	ParOps
Type	document level
Theory	by splitting paragraphs in two or combining two into one, we can mimic a target style's paragraph-length choices
Data sources	None
Technologies	Python, NLTK, Stanford part-of-speech tagger, list of transition words
Example	*[see below]*
Development	Foaad Khosmood. Implemented as Python class ParagraphOps.

ParOps is the only non-sentence level transform currently implemented in our system. It is also the only transform where possible document structure (titles and footers) is exploited. Structural markers have been used in several studies in the past [DeV01] [Zhe06] [Abb08], but we have avoided them in general in order for our analysis to be more widely applicable.

The transform can perform two operations: (1) Combine two paragraphs into one paragraph and (2) split a paragraph into two paragraphs preserving sentence integrity. The splitting operation uses a handcrafted scoring algorithm to determine where within a paragraph the split should occur.

This transform optionally takes in a parameter representing target paragraph lengths. These lengths are available from our own system's style marker gathering

phase and thus could be easily passed in. If the parameter is used, ParOps can determine which direction (split or combine) to go. Otherwise, ParOps just chooses one randomly.

The transform first tokenizes all paragraphs, sentences and words and furthermore part-of-speech tags the words before calling either combine() or split().

5.2.4.1 Combining paragraphs

Combining paragraphs is very simple. First the transform obtains lengths (in terms of sentences) of all the paragraphs. Then, it determines the two consecutive paragraphs that form the smallest paragraph, were they to be combined. The function can optionally ignore titles and footers for this process. Titles and footers are defined as single-sentence or single-line paragraphs near the top or the bottom of the document. Finally, the transform combines the two paragraphs and readjusts the document object to reflect it. Each call to combine() returns a document with at most one new paragraph formed out of two existing ones.

5.2.4.2 Splitting paragraphs

Splitting paragraphs with split() follows a similar process as combine. First, the largest single paragraph is found and designated for splitting. Now the question arises, where within the paragraph should the split occur? Given that some obvious hazards exist if the split were to happen in the wrong spot, we have developed an algorithm for scoring every sentence in the paragraph as a measure of how "split-able" the paragraph may be if a split were to happen such that a new paragraph would be formed beginning with that sentence.

Given a paragraph with N sentences, zero based sentence offsets 1 through N-1 are scored according to the following formula:

$$Score(n) = CenterScore(n) + ConjunctionScore(n) + TransitionScore(n) + Pronouns(n)$$

Table 17. Sentence scoring terms for *ParOps* split() function

Term	Formula	Explanation
$CenterScore(n)$	$= \frac{1.0}{1.0 + \lvert 0.5 * (N-1) - n \rvert}$	The closer the sentence is to the middle of the paragraph (in terms of sentences), the higher the value.
$ConjunctionScore(n)$	$= \begin{cases} 0.0, & if\ first\ word\ is\ a\ \text{conjunction} \\ 1.0, & otherwise \end{cases}$	First word having been POS-tagged as conjunction causes the sentence to score lower.
$TransitionScore(n)$	$= \begin{cases} 0.0, & if\ first\ word\ is\ a\ transition \\ 1.0, & otherwise \end{cases}$	First word's presence in a list of transition words causes the sentence to score lower.
$Pronouns(n)$	$= \frac{P(n)}{L(n)}$	P(n) is total number of words tagged "PRP" (pronoun) in sentence n, and L(n) returns the total length in words of sentence n. The higher the pronoun count, the more likely the sentence is semantically tied to previous sentences and hence a bad candidate for splitting.

The highest scoring sentence becomes the first sentence of the new paragraph with the previous paragraph terminating immediately prior to it. In case of tie, the sentence closest to the center is chosen. Part-of-speech tagging is necessary in order to

determine the second and fourth terms having to do with conjunctions and pronoun counts respectively.

5.2.5 PHRASE REPLACEMENT TRANSFORM

Table 18. Phrase replacement transform

Transform	Phrase
Type	sentence level
Theory	replacing phrases with their same-part-of-speech and same-sense synonyms should yield a semantically equivalent but different sentence
Data sources	WordNet, Microsoft Bing N-Gram database
Technologies	Python, NLTK, ConceptNet, Stanford part-of-speech tagger
Example	*I wanted to be with you alone => I desired to be with you only.*
Development	Foaad Khosmood. Implemented as Python class phraseXform.

The Phrase replacement transforms (Phrase) systematically examines a sentence for phrases (defined as consecutive words, or just one word) for which there may be a synonym available from the WordNet corpus. One or more response phrases are generated by the transform, each containing one valid substitution. The substitutions themselves are further validated internally by using contextual occurrence likelihood from the Bing N-Gram service. Thus if two synonyms for the same word are being considered, the one that occurs more frequently in the same context (as measured by Bing search engine) is the one used in the response.

These responses are passed on to the calling object. In this project they ultimately become part of valid replacement candidate pool.

As with the ParOps transform, Phrase can also accept an optional input, all words of the target corpus. If provided, Phrase will privilege the use of synonyms that also occur in the target corpus.

5.2.5.1 WordNet synsets

The WordNet corpus is organized into groups of synonyms of the same linguistic sense called "synsets" which in turn have various surface words, called lemmas, associated with them. If the correct synset for a word is found, other lemmas within the same set, subject to the same inflection rules, are true synonyms and are generally considered to be good substitutions for the original word.

Synsets also have part-of-speech (POS) designations. WordNet implements 4 POS categories: Nouns, verbs, adjectives and adverbs. Thus, all replacement activity of this transform is confined to those four categories. A string can be looked up in the WordNet corpus by passing it as a parameter to a lookup function. If the string exists a set of synsets are returned. One can further specify a supported POS to narrow down the search.

5.2.5.2 Word sense disambiguation (WSD)[1]

For a great many of the English language words and phrases, the returned POS-filtered result still consists of multiple possible synsets. This is because there are many words with multiple senses in English. The general problem of WSD is still open. The

[1] See [Mic05] or http://en.wikipedia.org/wiki/Word-sense_disambiguation for a general description

most-common-sense of each word is known to WordNet and that, in a general context, is measured to be about 70% accurate, which is considered the baseline for WSD algorithms. Many algorithms exist that can increase the accuracy slightly but the best perform generally still shy of 78-80% accuracy. We use a method described by [Mic05] for our WSD based on Jian and Conrath (jcn) similarity measure [Jia97].

5.2.5.3 Transform design

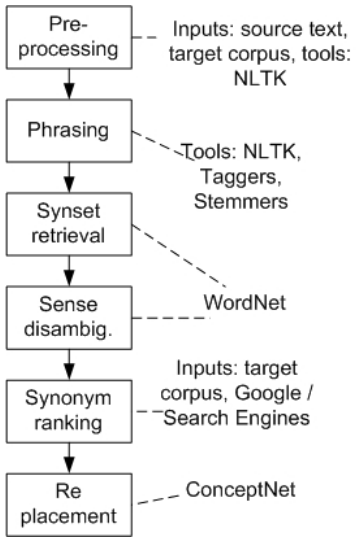

Figure 15. Phrase replacement algorithm steps and tools

A. **Preprocessing:** Prepares the sentence for operations, specifically this step performs:

   a. Some standard contraction expansions such as “it’s” to “it is”

   b. Word and sentence (if needed) tokenization and segmentation [Pal00]

   c. Part of speech tagging (using Stanford Parser)

B. **Phrasing:** The sentence is divided into a sequence of non-overlapping sentence fragments separated by punctuation marks including commas, colons, semicolons and quotation marks. Within these fragments, the largest phrase is searched for and found. A phrase, for our purposes can be a one, two, three, four or five word sequence in the same sentence fragment. Therefore a word is considered a phrase of length one. Each phrase will be examined for replacements from WordNet. In forming the phrase chunks, preference is given to the largest phrase that has a synonym in WordNet. Furthermore, the algorithm greedily selects phrases from left to right. Any phrase that is either shorter than the global minimum phrase length, or is part of an exclusion list, or is tagged as a proper noun is skipped over as part of a candidate for replacement. Phrasing is not a perfect process and major problems exist with recognizing named entities or multiword expressions, and hyphenated terms [Pal00].

C. **Synset retrieval:** Each remaining candidate is looked up for one or more synonyms in WordNet where synonyms exists in groups called synsets. Single words are lemmatized and have their POS tag also included in the WordNet lookup for further accuracy, i.e. only synsets with the identical tags are returned as candidates. In addition to words in the same synset, synsets of hypernyms are also retrieved. All replacement candidate synsets for a single phrase are placed in a list called *options*.

D. **Sense disambiguation:** For all nouns and verbs, a jcn distance-based disambiguation algorithm [Mic05] with a maximum context window of 5 is run which narrows down the list of eligible synset candidates in *options*. The top scoring jcn synset and any synset within 0.25 standard deviations of the top is returned. The pre-designated "most common sense" which is known to WordNet is also returned.

E. Synonym ranking: Every synonym lemma of the chosen synsets is scored based on the hit rate of its corresponding original lexeme in an online search engine (Bing). The joint probability of occurrence of the said lexeme within a bigram or trigram is obtained from Microsoft and weighed as factor in ranking the lemma. If this is a style shifting operation (as opposed to style obfuscation) and the optional word list from the style target is available, that information is used as well (i.e. a co-occurrence vector [Dag00] of bigrams and trigrams in the target corpus). The target statistical information, when available, is weighed higher than the search engine information. Both of these operations work on the principle of Distributional Hypothesis [Har68], which states that words occurring in the same context are likely to have similar meaning. At the end of this process, a single lemma should emerge as the most ideal replacement candidate.

F. **Replacement**: The lemma with the highest score is chosen to replace the original phrase. The replacement phrase is constructed by combining the

replacement lemma with the original phrase residue (all the non-lemmatized, inflectionary parts of the phrase). For example, the word "utilized" is lemmatized as "utilize" with the "ed" as residue indicating past tense. We reference "utilize" as a verb in WordNet and the word "use" is chosen among the synonym lemmas. Combining "use" and "ed" will produce "used" which is the actual replacement phrase. Non-standard inflections such as "went" and "attorneys general" are supported. Problems still exist in distinguishing between inflectional and agglutinative English words [Pal00]. For lemmatization and recombination operations, we used MIT ConceptNet tools [Con09].

5.2.5.4 Style obfuscation using Phrase

For this experiment, we used AAAC corpus, specifically the thirteen documents of problem A (acknowledged to be the most difficult). For the initial part-of-speech (POS) tagging, we used the Stanford log-linear POS tagger with the left-3-words algorithm that has performed with 96.97% on the Wall Street Journal data set. For synset lookups, we used the WordNet interface from NLTK. For lemma operations including lemma/residue split and combine functions we used the natural language tools from MIT ConceptNet. The transform itself was implemented as a python program.

The original and transformed samples were evaluated using JGAAP. The settings for running JGAAP (version 4.0) were the following:

1. Load AAAC problem A defaults, adding the transformed sample as a new document with an unknown author

2. Apply “Strip punctuation” to all documents in canonization step
3. Use “words” as event set
4. Use “RN Cross Entropy” as distance measure

We demonstrate that a stylistic shift is possible using phrase replacements. We process all 13 documents of problem A with the phrase replacement transform above. The before and after authorship attribution results are shown in the following table. The RNCE is the best method that is supported within JGAAP for this particular problem. Nine of the original thirteen problems were classified correctly to begin with. With their classifications shown before and after using the same exact black-box classifier, we can see that six documents’ styles were shifted away from what they were to begin with.

Table 19. AAAC problem A style obfuscation

Problem A Sample	True Attribution (Author #)	RNCE on original	RNCE on transformed	Stylistic shift
Sample1	3	12	12	no
Sample2	13	2	13	yes
Sample3	11	10	10	no
Sample4	7	4	4	no
Sample5	10	10	10	no
Sample6	12	12	13	yes
Sample7	8	8	2	yes
Sample8	1	1	9	yes
Sample9	5	5	5	no
Sample10	4	4	2	yes
Sample11	6	6	6	no
Sample12	2	2	9	yes
Sample13	9	9	2	yes

5.2.5.5 Targeted style transformation using Phrase

We would like to examine directed, i.e. biased, stylistic transforms whereby the phrase replacement algorithm specifically tries to bring the style closer to a particular target. We select an arbitrary sample among the sample texts that JGAAP/RN Cross Entropy can classify correctly. This was sample 6, which is written by Author 12. We also select another author, (Author 13) to serve as target corpus. The training documents labeled as author 13 were combined and word frequency statistics extracted as per step 1 of the phrase replacement transform. The Cross Entropy algorithm, which performed the best on problem A, compares every single training document with every test document and generates a distance with lower values indicating closer stylistic match.

Table 20. Targeted style transformation with AAAC corpus problem A

Document	RNCE dist. with A12	RNCE dist. with A13	Attribution
Sample6-original	**468.742** 478.497 4160.666	492.493 492.653 3475.155	Author 12
Sample6-transformed	464.770 480.477 4088.701	**462.662** 472.736 3331.260	Author 13

The two samples are "sample6" (original test document 6 attributed correctly to Author 12) and "sample6-transformed". The 3 numbers in the cells indicate the RN Cross Entropy distances between the sample and each of the training documents. The

algorithm selects the single lowest overall distance in all binary comparisons for attribution. These numbers are boldfaced in the table.

5.2.5.6 Evaluation of phrase replacement experiments

The evaluation for this kind of transformation problem is invariably two fold. First we must confirm that a shift in style has taken place, as detectable by some objective critic. Second, we must ensure that the transformed text is actually coherent and meaning preserving.

The first evaluation is accomplished using JGAAP. We treat JGAAP and the RNCE algorithm as an objective critic and authorship classifier, in fact the single best performing classifier in JGAAP for the given problem. This classifier, for example, has determined that our transformed text is closer to Author 13, than to Author 12.

The second part of the evaluation is rather difficult and subjective. We do not have a precise standard for evaluating the accuracy of these sentences. The original student essays themselves contain some mistakes and awkward language. In addition, for issues of delineating between "style" and "meaning" there is no firm community standard and definition. Some researchers, for example, consider almost any possible rephrasing to be a change of meaning.

We assume a dualistic approach to language and consider a single message to be communicable in a variety of styles, similar to the definition presented in [Wal80] which considers style an "option". We adopt a standard for evaluating the resulting sentences

inspired by automatic translation literature. Accordingly we find that the transformed sentences fall into three categories: “correct,” “passable,” and “not correct”. Our own examination indicates the majority of sentences are passable, but future work must involve human critics to make this determination.

For example the following sentence from the sample 6 text discussed above, is considered a good transformation. (The “=>” delineates before and after text).

> *Work provides more than mere [sustenance => nourishment] however , more importantly it provides an individual with [purpose => aim] in life .*

This one indicates a bad transformation, introducing some false or ungrammatical substitutions such as “water line” for “watermark” and “in one case” for “once”.

> *These [tasks => projects] have a [relatively => comparatively] low [watermark => water line] and are clearly delineated [once => in one case] we [achieve => accomplish] them , so [basically => essentially] [once => in one case] we have reached a certain [level => degree] in society.*

5.2.6 SIMPLIFY TRANSFORM (SIDDHARTHAN)

Table 21. Simplify transform

Transform	Simplify
Type	sentence level
Theory	The algorithm looks for certain conjunctions and appositions as an opportunity to split the sentence into two functioning sentences and thus reduce the complexity of the original long sentence.
Data sources	Built-in training models for Stanford tagger.
Technologies	Stanford parser, Perl, Morphological analyzer from RASP toolkit, Python wrapper
Example	*I love Luna who is very cute and who is my pet. => I love Luna. Luna is very cute. Luna is my Pet.*
Development	Advaith Siddharthan [Sid10] with slight rule modifications and Python integration by Foaad Khosmood.

This transform is based on the work of Advaith Siddharthan from his 2010 paper "Complex Lexico-Syntactic Reformulation of Sentences using Typed Dependency Representations" [Sid10]. The authors have developed and coded a number of handcrafted rules operating on typed dependency tree representations (provided by the Stanford Parser) to re-write eligible sentences. The system was tested on a corpus of 144 complex sentences in four categories of rewrite rules and it performed above 72% F-measure in all categories. We refer the reader to the original paper for more details of the algorithm and performance evaluation.

5.2.6.1 Rules system

The system works by first getting a typed dependency tree from the parser, and then manipulating the tree using rules. Finally, the resulting text is generated using only

the word ordering of the original parse with some morphological operations on verbs and nouns. Thus the rules only apply directly to the typed dependency parse trees. The system tries to resolve the given parse tree against a pattern specified in the rules in the "<DELETE>" section. If successful, the pattern is deleted and a modified one is inserted from the <INSERT> section. The tree then is passed on to a simple realizer that follows the original word order to generate the final text. If no feature exists to link two clauses (such as the one that may have been deleted by the rule), then clauses become their own independent sentences.

```
<RULE>
<TYPE>
conj_and_clause
</TYPE>

<DELETE>
conj_and(??X0, ??X1)
nsubj(??X0, ??X2)
nsubj(??X1, ??X3)
</DELETE>

<INSERT>
nsubj(??X0, ??X2)
nsubj(??X1, ??X3)
amod(??X1, and-0)
</INSERT>

<ORDER>
</ORDER>

</RULE>
```

Figure 16. Sample rule from *Simplify*

In the figure 15 below, a conceptual demonstration is shown for how the sentence's typed dependency tree is manipulated by the system. The tree to the left of

the arrow is generated from the parser, however the picture on the right side of the arrow, is purely conceptual and does not exist in tree form in the work of Siddharthan. Rather than construct the tree and then proceed to surface realization, Siddharthan jumps straight to surface realization by using original order of words as modified by the rule application.

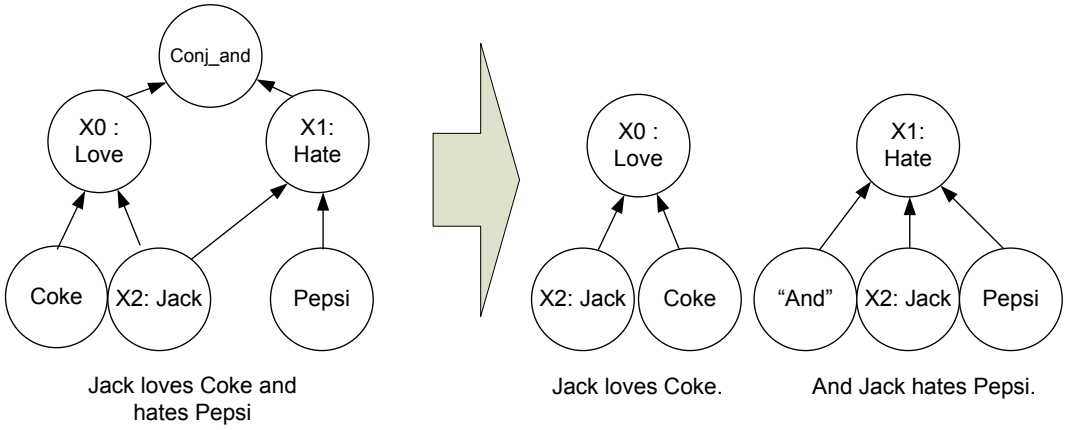

Figure 17. Sample conceptual operation by *Simplify*

5.2.6.2 Adaptation for use in this work

We adapted the original Perl-based code supplied by the authors for use in our system. Using a Python wrapper, we created a new class that communicated to the original code and captured its output directly into Python objects, as opposed to files, as originally designed. We also added rules to make conjunction based sentence divisions easier, being that our system has an additional level of syntax checking after transform-level of generation. Specifically, our new rules cover sentences with more than one

conjunction sharing the same subject. For example: “I came and I saw and I conquered,” or “We don’t allow eating or drinking or smoking.”

5.2.6.3 Testing for robustness and penetration using the King James Bible

Since the experiment on the Siddharthan system was of 144 specialized sentences, we test the system with the basic rule set for robustness and penetration (likelihood of application). For each sentence, if the system did not return anything, halted or became unstable, the robustness value of zero would be given, otherwise, one. If the system did return a sentence and the sentence was actually a typographical modification of the original (including punctuation mark modification), then a penetration value of zero would be recorded for that sentence. If the return value was the same as the input value, penetration would be zero.

We tested the system for both of the above criteria. We used the King James Bible corpus, from project Gutenberg as bundled with NLTK. A total of 50,803 sentences were used in the experiment. Out of those 19,635 were successfully modified and 3920 caused system failures.

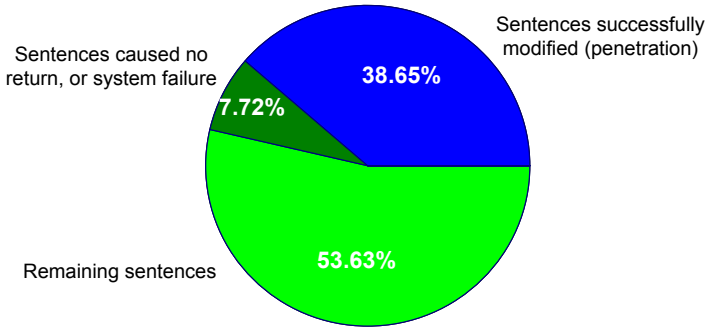

Figure 18. King James Bible simplify results by percentage

We note that there is no statement about accuracy made in these results. The roughly 20,000 sentences must be evaluated for grammaticality and logic in order to determine accuracy. However, we can at least arrive at a rough idea of an upper bound for accuracy given 38.65% penetration.

Some example results with self-evaluation in accuracy are given below. Accuracy is judged by the translation standard of ("good", "bad" and "acceptable").

Table 22. Example *Simplify* results

Before	After	Accuracy
Reward her even as she rewarded you , and double unto her double according to her works.	Reward her. This is even as she rewarded you. double unto. Unto her double according to to her works.	bad
In the cup which she hath filled fill to her double.	The cup which she hath. In the cup filled fill to her double.	acceptable
How much she hath glorified herself , and lived deliciously , so much torment and sorrow give her.	This is how much she hath. Glorified herself, and lived deliciously, so much torment and sorrow give her.	bad
For she saith in her heart, I sit a queen, and am no widow, and shall see no sorrow.	This is for she saith in her heart. I sit a queen, and am no widow. I	good

	shall see no sorrow.	
For strong is the Lord God who judgeth her.	For strong is the Lord God. The Lord God judgeth her.	good

5.2.7 TRANSLATION TRANSFORM

Table 23. Translation transform

Transform	Translation
Type	sentence level
Theory	Automatic translation tools usually produce valid but slightly modified sentences. A translation from English into another language and back to English could provide a valid semantic variation of the original with sufficient quality.
Data sources	Microsoft Translator, Yahoo Babelfish Translator, Tatoeba English sentence database.
Technologies	Python, NLTK
Example	*I like to play Tennis. => I enjoy playing Tennis.*
Development	Foaad Khosmood. Implemented as Python class TranslationXform

The translation transform uses third party, automatic translation tools to generate a similar version of the same sentence. Two main flavors of this transform have been explored. First is primarily based on translation into one language and then back again to English. The second is based on translation into several languages serially and then back to English at the very end of the chain. We use the first version in our translation transform, but the second is also used for comparison purposes.

5.2.7.1 Transform design

The Translation transform based on a "to and from" translation paradigm, consists a series of these translations in order until one is found that is typographically different than the original English. The list of languages is prioritized in order of likely accuracy, derived from an experiment by the authors, discussed later in this chapter.

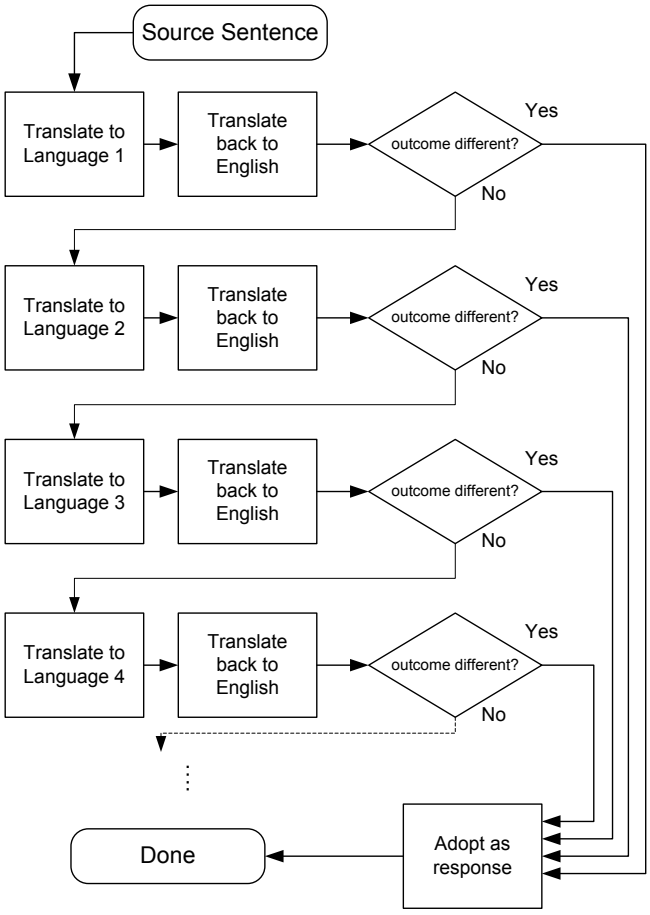

Figure 19. Translation transform design

5.2.7.2 Deriving a list of best suited languages for pivot translation

In this section, we experiment with the Microsoft Translator and derive an ordered set of languages to use in the Translation transform. We use 31 random sentences from the Tatoeba English corpus. These sentences are then translated into each of the 32 (at the time of experiment) API supported languages and back again to English. The results are evaluated by the authors on a scale of 0-3.

Table 24. Microsoft Translator supported language codes (ISO 639-2)

Code	Language	Code	Language	Code	Language
ar	Arabic	he	Hebrew	ro	Romanian
bg	Bulgarian	hu	Hungarian	ru	Russian
cs	Czech	id	Indonesian	sk	Slovak
da	Danish	it	Italian	sl	Slovenian
nl	Dutch	ja	Japanese	es	Spanish
et	Estonian	ko	Korean	sv	Swedish
fi	Finnish	lv	Latvian	th	Thai
fr	French	lt	Lithuanian	tr	Turkish
de	German	no	Norwegian	uk	Ukrainian
el	Greek	pl	Polish	vi	Vietnamese
ht	Haitian	pt	Portuguese		

The coding scheme roughly reflects a standard human translation coding scheme where three categories are available: “Good”, “Acceptable” and “Bad”. To this we add one more category that would be “terrible” or “unresponsive” to cover cases of machine problems or software bugs. The criteria for this category are: There must be no response returned from the routine, or a very poor/incompressible response returned or machine error returned.

Table 25. Translation coding scheme

Code	Criteria
3	Sentence is modified, preserves original meaning and original level of grammaticality.
2	Sentence is unmodified, or barely modified with very minor changes, or modified and the result is “acceptable” but not great.
1	Sentence is modified and the result is not acceptable.
0	No response, error or incomprehensible response

Each language was in effect evaluated 31 times with different sentences. The average scores for each language are given in Figure 20 below. At this point, we begin the process of sub-selecting languages to include in the Translation transform.

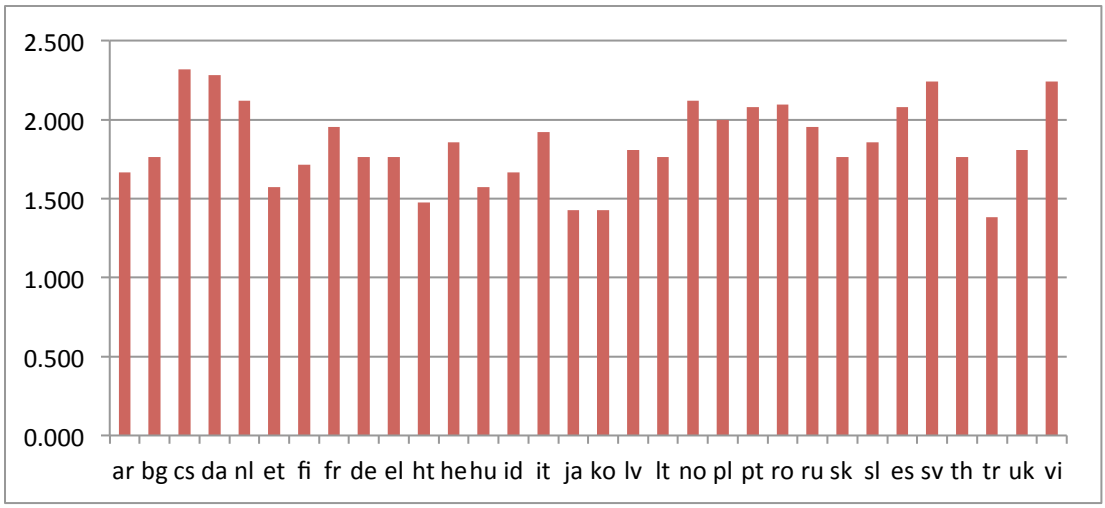

Figure 20. Average score of Microsoft Translator languages

We are most concerned about correctness and accuracy (as style is entirely dependent on the target we are trying to achieve). Therefore, as a first step, we eliminate any language that produced a "0"-score from the contention. A chart of "0"-score performances is provided below.

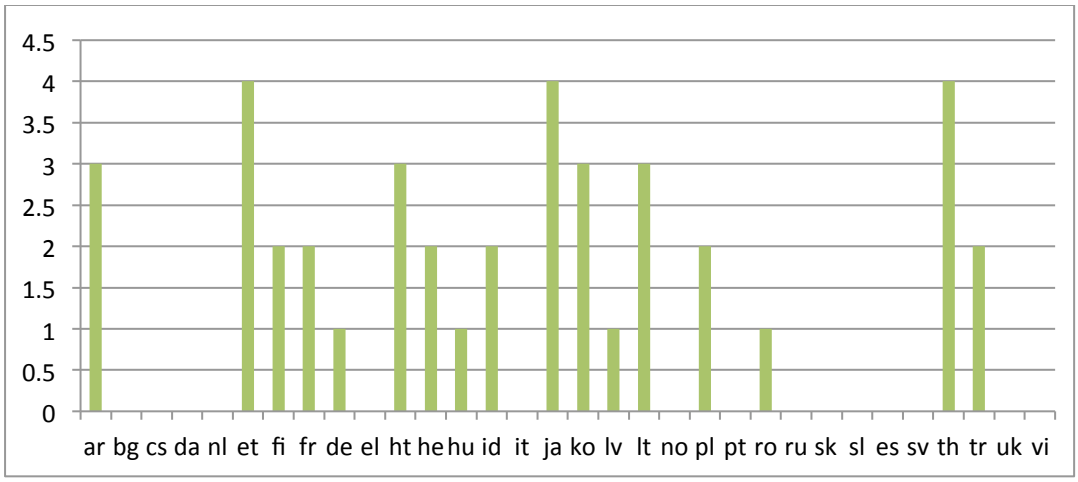

Figure 21. Instances of "0"-scores per language

Noting that the theoretical average score of 2.0 represents the same sentence unmodified or one of "acceptable" quality, we adopt this level as the floor of our next selection criteria. We eliminate all languages that do not have an average performance score of at least 2.0. Ten languages scored at 2.0 or above level: Czech, Danish, Dutch, Norwegian, Polish, Portuguese, Romanian, Spanish, Swedish, and Vietnamese. Of these Polish and Romanian contained "0"-scored sentences. Thus we are left with eight languages.

Finally we define a "goodness ratio" which consists of the number of "3"-scored sentences divided by the number of "1"-scored sentences. We can restate this ratio as "how many excellent responses are likely to occur per one bad response?" With this, we list the remaining languages and rank them according to goodness ratio.

Table 26. Goodness ratio and final ranking of languages

Language	1-scores	3-scores	Goodness ratio	Final ranking
Czech	2	11	5.500	1
Danish	2	8	4.000	3
Dutch	3	7	2.333	6
Norwegian	3	8	2.667	5
Portuguese	3	5	1.667	7
Spanish	6	8	1.333	8
Swedish	2	10	5.000	2
Vietnamese	2	8	4.000	4

Spanish has the highest number of "1"-scores and is accordingly ranked eighth or last. However, this is not such a high concern because at each ranking, only if the language to-and-back translation produces the same output as the input, the system actually moves to the next ranking language. Thus the probability of Spanish actually being used is quite low. Most of the translations in the Translation transform can be said to be by the first few languages in ranking order.

5.3 Comparison and relative performance of sentence level transforms

In this section, we compare the performance of different sentence level transforms discussed in Section 5.2. When comparing transforms two distinct measures must be considered separately:

- How likely is the transform to be capable of modifying a random sentence?
- How accurate is the transform in terms of how grammatical and meaning-preserving the response?

Some transforms are highly accurate in that they produce valid sentence reformulations. But if they only apply to a small number of sentences then their accuracy is no longer as important. There can be dozens or even hundreds of very accurate transforms but they are very unlikely to apply to a random English sentence. On the other hand, we may have transforms that can operate on almost every sentence, but their accuracy is low and therefore applying them will result in many undesirable mistakes.

5.3.1 REDEFINING "PRECISION" AND "RECALL"

We adopt the Information Retrieval (IR) concepts precision and recall for the purpose of reporting performance of transforms. The concept of a system's recall in IR is defined as "the proportion of relevant material actually retrieved in answer to a search request" [Rij79]. Precision is defined as "the proportion of retrieved material that is actually relevant." For our purposes, we use these definitions:

- "Recall" is the proportion of the text (sentences in this case) for which the transform routine has returned a valid result other than the input.
- "Precision" is a measure of accuracy for the intended transformation defined as the proportion of the transformed sentences that have preserved the level of semantic meaning and the grammaticality of the original sentence to an acceptable level.

The application of precision and recall to text manipulation methods also has some precedence. In [Sid10], the authors used precision and recall to evaluate sentence re-generation results in a very similar way using nearly identical definitions.

5.3.2 Evaluation of transforms using 50 random English sentences

In this section we evaluate six sentence level transforms and three derivative translation-tours for comparison purposes. 50 sentences from the Tatoeba [Tat10] English language corpus are evaluated by each of the transforms. The resulting 381 transformed responses are evaluated by three human evaluators in two separate sessions. The criteria for the evaluations are exactly those specified in Table 25 above. A summary of the transforms used in this experiment are given in the following table.

Table 27. Transforms used in evaluation experiment

Short form	Transform	Notes
A2P	Active to Passive transform	
Diction	Diction transform	also includes 2 other data sources
Nodebox	Nodebox transform	
Phrase	Phrase replacement transform	
Simplify	Simplify transform (by Siddharthan)	
Translation	Translation transform (uses Microsoft Translation)	used as baseline against other author-developed transforms
Tour1	Translation tour with Spanish, French, German	3 commonly translated languages in the United States used for comparison purposes
Tour2	Translation tour with Danish, Portuguese, Swedish, Vietnamese	4 high performing BING languages chosen based on experiments in Section 5.2.7.2, for comparison purposes
TourR	Translation tour with four random	random tour for

	languages	comparison purposes

We derive the relevant precision and recall statistics as follows.

- "recall" is calculated by counting the number of sentences for which the transform had a response (other than the input), divided by the total number of sentences, 50.
- "precision" is the average of all the (0-3) scores given by human evaluators to any response of the transform, across all sentences.
- "F-measure" is defined as $F = \frac{2*(precision*recall)}{precision+recall}$

Table 28. Transform experiment evaluation results in terms of precision, recall and F-measure

	A2P	**Diction**	**Node box**	**Phrase**	**Simplify**	**Trans-lation**	**tour1**	**tour2**	**tour R**
recall	10%	4%	20%	100%	7.6%	74%	88%	78%	88%
precision	80%	100%	87%	82%	78%	83%	63%	70%	58%
F - measure	0.178	0.077	0.325	0.903	0.138	0.782	0.733	0.738	0.702

As can be seen in Figure 22, Phrase seems to be performing relatively high on both axes. Phrase also has the highest F-measure. In general the transforms can be divided into two groups: high penetration and low penetration. The high penetration group consisting of Phrase, Translation and the Tours all apply to 74% or more of general sentences. The low penetration group consisting of A2P, Diction, Nodebox and Simplify understandably apply to a lesser portion of general sentences. Most of these transforms have handcrafted or narrowly defined rules that limit their applicability but at the same time boost their precision performance. Diction even reached 100% in precision meaning that all Diction-response sentences scored a "3" by all human evaluators.

Using the Translation as a baseline, we find that it is generally high penetration, which is consistent with the general web-translation utility that it derives from. However, its precision is lower than Diction and Nodebox and almost at the same level as Phrase. In general successive translation to more than one language has a predictable degradation in quality resulting in tour1 and tourR having the lowest precision in the group. Unsurprisingly, tour2 has a higher precision due its usage of higher performing languages, as derived in Section 5.2.7.

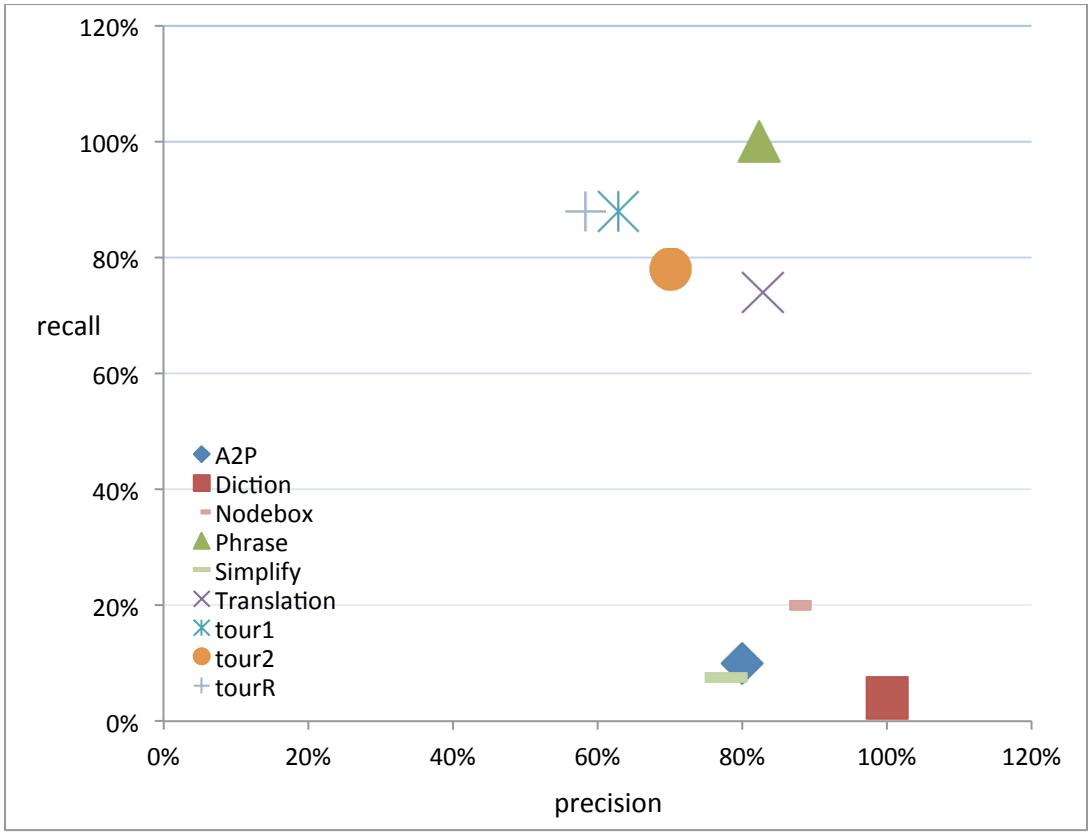

Figure 22. Precision and recall in 9 transform modules

5.4 COMBINING THE POWER OF ALL TRANSFORMS OR "HOW TO PICK THE BEST SENTENCE"

In this section, we answer the question "How do we pick the best sentence." The question is important for several reasons. First, finding a method to pick the best of many choices allows us to actually combine the power of all the transforms. Without such a method, we wouldn't really know how to pick the best response among the many (perhaps dozens) that could be produced by all the transforms. Second, while many transformed sentences may be technically correct in both meaning and grammar, they are nevertheless undesirable. For example "I locomote to school with a two wheeled mechanical device," is technically a valid transformation for "I bike to school." However, there are very few circumstances that would require the use of the former rather than the latter. If substituted during a real transformation exercise, the longer sentence will generally add to the awkwardness of the text, perhaps to a distracting level. Thus in addition to avoiding possible ungrammatical responses, we must somehow also avoid the "unnatural" or "awkward" ones as much as possible.

5.4.1 STATISTICAL FEATURES OF GOOD SENTENCES

For this project, we use two categories of metrics that we can derive from a general sentence that we could later use to evaluate it. These are 1) Internet search hit rates and 2) Large English corpus word co-occurrences.

5.4.1.1 Using BING search results to validate sentences

The idea behind Internet search hit rates is simple. If one group of consecutive words have higher hit rates on the Internet compared to another, then that group must be more commonly used and therefore likely to be more natural sounding. For example, if we compare "I saw an owl" to "I saw a owl", the former will have more hit rates and thus we can avoid a grammatical mistake by simply looking up a string on the Internet.

The Microsoft Web N-Gram service [MSN11] provides a perfect opportunity to evaluate sentences using search engine hit rates. The service provides Microsoft BING search engine-crawled statistics reflecting the state of the World Wide Web as of April 10, 2010. The service accepts input in form of a sequence of 2, 3, 4 or 5 words. It then returns the log of the joint probability of those words occurring in that sequence. We concentrate on word bigrams and trigrams in main text of web pages. This is mostly due to the fact that smaller sentences cannot be evaluated using the 4 or 5-gram models. The service ignores punctuations and word capitalization.

For both bigram and trigram models, we can derive an average, a minimum and a maximum value per sentence. The average is calculated by taking the average return value of every n-gram present in the sentence. The lowest n-gram log-probability encountered would be the minimum and the highest the maximum. We use both minimum and average for our purposes and ignore the maximum. As the owl sentence demonstrated above, often the problem is one misplaced word or combination. Thus a low minimum n-gram value would serve to isolate that misplaced construction (i.e. "a owl") and is very useful. However, the maximum, which simply reflects the most

popular n-gram in the sentence, does nothing to indicate potential problems elsewhere in the sentence.

Table 29. Web N-Gram statistics per sentence

Model	Value	Method of calculation
word bigrams / webpage body	average	average of all roving window bigrams in the sentence
word bigrams / webpage body	minimum	minimum bigram probability encountered
word trigrams / webpage body	average	average of all roving window trigrams in the sentence
word trigrams / webpage body	minimum	minimum trigram probability encountered

5.4.1.2 Using the Brown corpus to validate sentences

In a similar fashion to Web N-Grams, we use the Brown corpus [Fra64] to get more features from sentences that we can correlate to their appropriateness. The corpus is large and topically diverse. Sentences that appear in it can be assumed to be correct and natural sounding. The idea here is to use a word co-occurrence vector [Dag00] with sentence-length windows size. Co-occurrence of words within the same sentence in the Brown corpus is considered evidence of relatedness of those words. Thus if a sentence we are examining has the same co-occurring words, we can take that as evidence for its correctness and naturalness. For example, given the following two sentences:

- I deposited my check at our friendly neighborhood bank.
- I deposited my dog at our friendly neighborhood bank.

We should be able to eliminate the second and choose the first. It is likely that word combinations of "check" and "bank" appear somewhere in the same sentence in the Brown corpus, but "dog" and "bank" probably much less often if ever. We also note that the N-Gram data would not help make the decision here, as the two words are too far apart for the bigram and trigram data to be used to evaluate them.

We can derive similar statistics using this co-occurrence concept. These are outlined below.

Table 30. Brown corpus co-occurrence statistics per sentence

Value	Elaboration and Calculation
BC-avg	Brown co-occurrence average: The concurrence hit rate (number of sentences in Brown that contain both words) for every combination of 2 words in the sentence.
BC-bin-avg	Brown co-occurrence binary average: Same as above, but instead of actual hit rates a 0 or 1 is returned. A 1 means at least one sentence contains the word combination.
BC-min-drop0	Brown co-occurrence minimum (above 0): Returns the lowest non-Zero hit rate from any combination of words in the sentence.
BC-sum	Brown co-occurrence sum: The sum of all hit rates of all word pairs in the sentence.
BC-max	Brown co-occurrence maximum: The biggest numerical hit rate among all the word pairs.

5.4.2 CORRELATING THE STATISTICAL FEATURES OF SENTENCES TO HUMAN EVALUATIONS

Using the same human annotated sentence corpus from 5.3.2, we derive correlations between average normalized sentence features and human evaluations (and also among the features themselves.) Pearson correlation is used as the standard

in statistical correlation. However, we also derived the Spearman correlations where the main difference is that relative values between ranked items are discounted.

Human evaluations are produced in two formats. First is average of all three evaluators (user-avg) and second is a "voted" score which is quantized to be at values of 0, 1, 2 or 3, normalized. In case of a tie, the average was used in the voted evaluation (user-vote) as well. Nine statistical measures are evaluated.

Table 31. Pearson correlations

	2g-min	**2g-avg**	**3g-min**	**3g-avg**	**bc-avg**	**bc-bin-avg**	**bc-sum**	**bc-min-0**	**bic-max**	**user-avg**
2g-avg	0.6815									
3g-min	0.7974	0.4729								
3g-avg	0.5391	0.6637	0.6634							
bc-avg	0.0635	0.316	0.0111	0.1658						
bc-bin-avg	-0.024	0.2349	-0.1805	0.1545	0.3207					
bc-sum	0.0231	0.3011	-0.0554	0.1542	0.9316	0.3686				
bic-min*	0.0795	0.0811	0.0992	0.0225	0.0691	-0.0617	-0.0166			
bc-max	0.0533	0.3103	-0.0211	0.1638	0.9497	0.3628	0.9495	0.0323		
user-avg	0.0934	0.1939	0.0548	0.1493	0.1104	0.123	0.115	-0.0517	0.0932	
user-vote	0.0905	0.1986	0.0401	0.1356	0.0975	0.115	0.1203	-0.0826	0.0877	0.945

Table 32. Spearman correlations

	2g-min	2g-avg	3g-min	3g-avg	bc-avg	bc-bin-avg	bc-sum	bc-min-0	bic-max	user-avg
2g-avg	0.6928									
3g-min	0.7714	0.5135								
3g-avg	0.5702	0.6499	0.7546							
bc-avg	0.0075	0.3717	-0.0944	0.1737						
bc-bin-avg	-0.0944	0.17	-0.2398	0.0122	0.5893					
bc-sum	-0.0335	0.3384	-0.1454	0.1422	0.9788	0.635				
bic-min-0	0.0769	0.3883	0.2269	0.2438	0.358	0.0586	0.2783			
bc-max	-0.001	0.3417	-0.1101	0.1582	0.9884	0.5904	0.9818	0.3308		
user-avg	0.1138	0.1781	0.0841	0.1649	0.1247	0.1414	0.136	0.0827	0.0119	
user-vote	0.1556	*0.2375*	0.1179	*0.1925*	*0.1762*	*0.1760*	0.1825	0.1375	*0.1710*	0.939

The strongest observable human-to-feature correlations are those in the Spearman table with the single strongest value being the bigram averages from Microsoft Web N-Gram data. We have italicized five of the nine sentence features as exhibiting the strongest correlations to user scores. These are: 2-gram and 3-gram averages from Microsoft N-Grams and Brown correlation average, binary average and maximum from the Brown data.

5.4.3 COMBINING FEATURES FOR AN OVERALL PREDICTION ALGORITHM

We experiment with three distinct methods of combining the top sentence level features (as derived in the previous section). Two of the methods are regression based and one is a decision tree based on correlation performance. To make it easier to

discuss these features in this section and the next, we will designate a function to stand for each of them.

Table 33. Sentence quality feature function designation

Sentence features	Function designation (input sentence)
Average word bigrams from Microsoft Web N-Grams	$F_1(S)$
Average word trigrams from Microsoft Web N-Grams	$F_2(S)$
Brown Correlations average word pair hits	$F_3(S)$
Brown Correlations binary average word pair hits	$F_4(S)$
Brown Correlations maximum word pair hits	$F_5(S)$

5.4.3.1 F1 formula: a decision-tree with tie breaking

This method is simple. It uses the Spearman correlation data to determine which of the feature values are most important and uses that to make a determination of which sentence should be scored higher. In case of a tie (which are frequent given that many sentences are very similar), the next highest feature value should be examined.

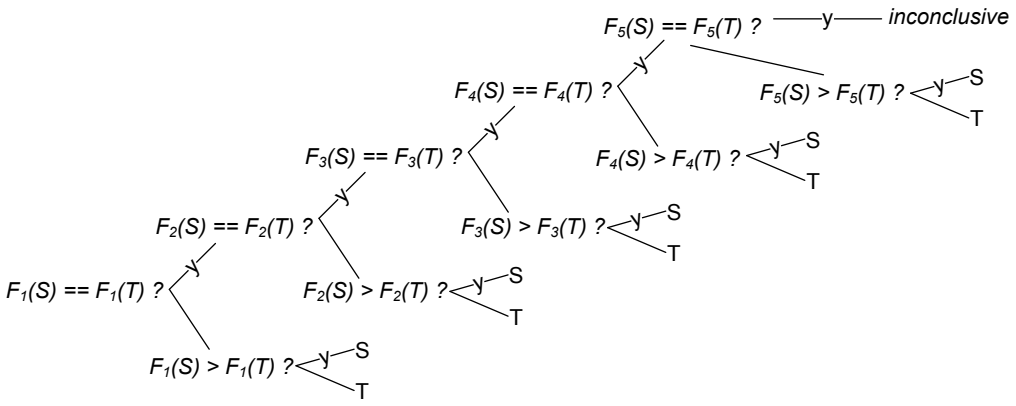

Figure 23. F1 "decision tree tie break" pair wise sentence comparison algorithm between sentences (S) and (T)

5.4.3.2 Linear regression among the five top features

The linear regression sets up the following formula for each of the human-scored training sentences and attempts to derive the coefficients that cause the outcome to be closest to the human scored value.

$$LR(S) = A_1F_1(S) + A_2F_2(S) + A_3F_3(S) + A_4F_4(S) + A_5F_5(S) + A_6$$

Where F_1 to F_5 denote the five chosen features from 5.4.2 and A is the coefficient matrix. We calculated the following values for the coefficient matrix.

Table 34. Linear regression coefficients

A_1	A_2	A_3	A_4	A_5	A_6
-4.47E-02	1.11E-01	-5.31E-01	7.19E-02	4.01E-01	-5.1E-04

5.4.3.3 Linear power set regression among the five top features

The linear power set regression is like linear regression except with additional terms and coefficients. The terms represent every combination of the five original terms multiplied together. For five terms this produces 31 different multiplicative combinations. With the additional linear offset term, this makes 32 coefficients that we derive with the training data.

$$LPSR(S) = A_1F_1(S) + A_2F_2(S) + A_3F_3(S) + A_4F_4(S) + A_5F_5(S) + A_6F_1(S)F_2(S)$$
$$+ A_7F_1(S)F_3(S) + A_8F_1(S)F_4(S) + A_9F_1(S)F_5(S) + A_{10}F_2(S)F_3(S)$$
$$+ A_{11}F_2(S)F_4(S) + A_{12}F_2(S)F_5(S) + A_{13}F_3(S)F_4(S) + A_{14}F_3(S)F_5(S)$$
$$+ A_{15}F_4(S)F_5(S) + A_{16}F_1(S)F_2(S)F_3(S) + A_{17}F_1(S)F_2(S)F_4(S)$$
$$+ A_{18}F_1(S)F_2(S)F_5(S) + A_{19}F_1(S)F_3(S)F_4(S) + A_{20}F_1(S)F_3(S)F_5(S)$$
$$+ A_{21}F_1(S)F_4(S)F_5(S) + A_{22}F_2(S)F_3(S)F_4(S) + A_{23}F_2(S)F_3(S)F_5(S)$$
$$+ A_{24}F_2(S)F_4(S)F_5(S) + A_{25}F_3(S)F_4(S)F_5(S)$$
$$+ A_{26}F_1(S)F_2(S)F_3(S)F_4(S) + A_{27}F_1(S)F_2(S)F_3(S)F_5(S)V$$
$$+ A_{28}F_1(S)F_2(S)F_4(S)F_5(S) + A_{29}F_1(S)F_3(S)F_4(S)F_5(S)$$
$$+ A_{30}F_2(S)F_3(S)F_4(S)F_5(S) + A_{31}F_1(S)F_2(S)F_3(S)F_4(S)F_5(S) + A_{32}$$

Table 35. Power set regression coefficients

A-coefficient	value	A-coefficient	value	A-coefficient	Value
1	-1.73E-02	12	-6.99E-02	23	1.08E-01
2	5.60E-02	13	-1.93E-01	24	1.55E-02
3	-4.83E-01	14	-2.44E-01	25	2.24E-02
4	1.14E-01	15	2.68E-02	26	-3.69E-02
5	2.98E-01	16	-2.01E-01	27	-1.27E-01
6	-3.62E-02	17	-9.85E-02	28	3.38E-02
7	2.29E-01	18	1.01E-01	29	3.70E-03
8	1.14E-02	19	-1.89E-02	30	1.46E-02
9	-2.21E-01	20	6.35E-02	31	-1.18E-03
10	2.16E-01	21	-1.91E-02	32	2.29E-01
11	1.43E-01	22	7.69E-02		

5.4.3.4 Evaluation of combination methods

We conducted an additional study between the three feature combination methods. We first applied them back to the original 50 sentences to see which one would have produced the best results. Out of the 50 sentences evaluated, the algorithms performed as follows:

Table 36. Comparison of feature combination methods

	F1	Linear	Linear Power Set
All derivations	13	9	10
Max-1 per group	10	9	10

When considering "Max-1" per group, meaning consider just one response from all the transforms, F1 and Linear power set are very competitive with regular linear regression method consistently behind. With this, we evaluated another 25 sentences and considered the results just between F1 and linear power set regression. Here, we decisively find that linear power set is the best method as indicated by the evidence.

390 transformed responses from the original 25 sentences were evaluated by human and run through F1 and linear regression. In 18 of the 25 instances, linear power set method was able to pick a sentence from the response group that was also rated highest by the human scorers. F1 was only able to accomplish this for 11.5 of the instances.

6.0 USER STUDY

We concluded a user study with 19 subjects in order to demonstrate correlations between our statistically derived model of language style, and style, as perceived by human readers. In this chapter, we present details of the user study protocols and questions as well as the results obtained.

Table 37. User study summary information

Name	Style or "Stylogic" user study, also called "Foaad's Dissertation User Study"
Dates	Saturday July 30, 2011 to Tuesday August 2, 2011
Format	Multiple choice questions requiring reading sentences and English language passages and 3 free-form feedback text boxes. Users conducted the survey over the web using the Survey Monkey service. All questions required a response except for two "additional feedback" free-form text boxes. A short phone interview (about 5 minutes) was held with users who completed the survey.
User Qualifications	Users were instructed to participate if they are "18 years or older and have good command of the English language."
Participants	22 users began the survey 19 users completed the entire survey
Duration	• Approximately 30-45 minutes for the web based portion • Approximately 5 minutes oral interview
Reward	$20.00 reward was offered to participants who completed the survey and subsequent phone interview.
Content	• 30 sentences evaluated on grammaticality and preference by user • 11 sentence pairs evaluated on meaning preservation • 2 Multi-paragraph passages evaluated on stylistic similarity to several text excerpts that were provided to the users • 2 Multi-paragraph passages evaluated for goodness of writing • 2 Multi-paragraph passages evaluated based on whether they were machine generated
IRB Status	An adults-only exemption from Institutional Review Board (IRB) was granted by UCSC Office of Research Compliance on July 15, 2011. UCSC IRB protocol #: HS1101710

6.1 USER STUDY DATES, FORMAT AND PARTICIPANTS

The study was announced on Saturday July 30, 2011 via email on a popular graduate student email alias at University of California Santa Cruz. Students were told this is a study about writing styles, and that heavy reading would be required. The length would be about 45 minutes but "some familiarity with Melville or Hemingway" may speed up the process. A link was provided to the Survey Monkey web-based service. Participants continued to take the survey until August 2, 2011. The author also invited participants using other venues such as email and personal encounters. The majority of the users, however, were UC Santa Cruz graduate students who responded to the general announcement.

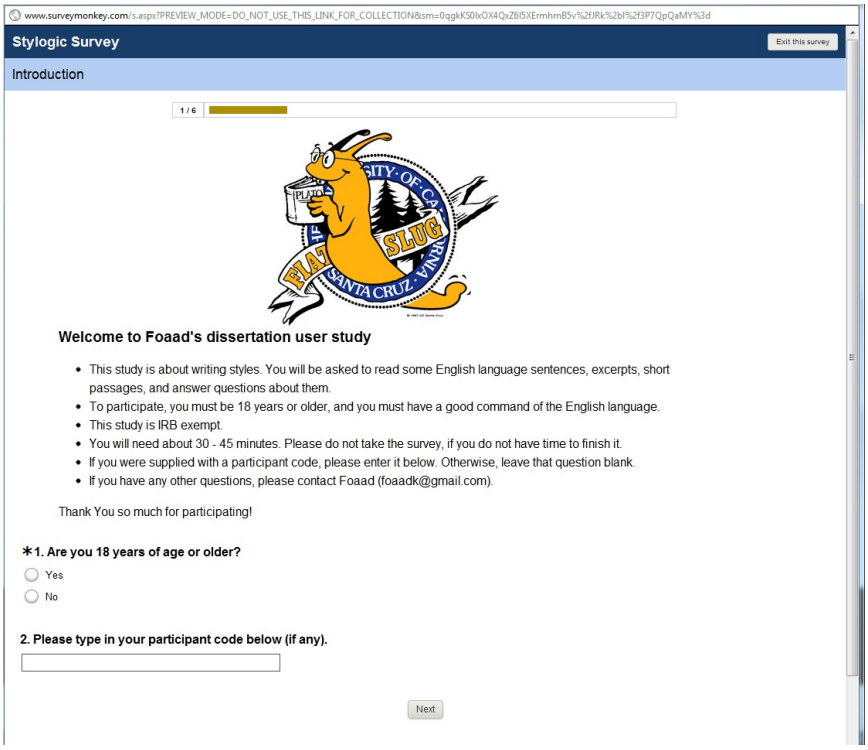

Figure 24. Survey Monkey web interface showing the first page of the survey

According to Survey Monkey statistics, 22 users began the survey and 19 users completed the entire survey. Three users began but did not complete any portion of the survey beyond the administrative questions. All results reflect the choices of the 19 users who completed the survey.

An additional 5-minute survey was conducted by phone with participants who finished the survey. Everyone was asked about general impressions, and additional insights into their decision-making.

6.2 USER STUDY QUESTIONS AND GOALS

In this section, we introduce the actual questions asked in the user study. For each question a statement of goals is provided, illuminating the reason for posting the questions and the hypotheses behind them. Questions 1 and 2 (displayed in Figure 24) are administrative questions and will not be covered any further.

6.2.1 SENTENCE EVALUATIONS (QUESTIONS 3 AND 4)

One hypothesis of our work, adopted from [Wal80] is that style is a particular selection of choices among the options available. In writing, this means that some distinct options of expressing the same meaning must exist that are roughly equivalent. Style, then becomes the array of choices made in favor of one set of options over others. A style processing system, we contend, must be able to produce these distinct options and then subsequently choose the ones most consistent with a target corpus.

The first part of that work is the production of different options for expressing the same original text. At a sentence level, that means different sentences saying the same thing but being lexicographically distinct. We thus have two hard criteria for production of alternate sentences given an original one: 1) That they make sense in English, and 2) That they preserve the meaning of the original. Beyond this, we also have a general preference for sentences that sound "natural" so that they would not be distracting to the reader and could match other human-generated sentences of a particular style.

Question 3 on the survey tests that sentences we generate are considered grammatical and natural by the users. Question 4 determines to what extent the generated sentences preserve the meaning of the original versions.

To have a successful style processing system, we expect an adequate level of grammatical sentence (re)generation and meaning preservation. Some degradation in quality and naturalness of writing is to be expected.

6.2.1.1 Obtaining sentences for evaluation

We first choose 15 original sentences from two separate corpora. These are:

- 8 random sentences from the AAAC corpus problem A consisting of student essays (but not any of the same essays used in subsequent survey questions)
- 7 random sentences from the Tatoeba English language sentence library

While we originally had planned on selecting all the sentences from the Tatoeba library, it became clear that the sentences there are generally shorter and less complex. The sentences from AAAC are about specific subjects and arguments. They are generally longer and more complex. In addition since we are using the AAAC corpus for the subsequent passage-level evaluation in the survey, we thought it appropriate to test the systems sentence level performance on the same quality data. All sentences were written by people and have arguable points of weakness in grammatical correctness and clarity.

The 15 original sentences were then transformed using a general un-targeted language obfuscation operation by our system resulting in 30 total sentences.

6.2.1.2 Individual sentence evaluations for grammar and naturalness (survey question 3)

For question 3, we displayed all 30 sentences in random order to the user. The user is not told at this point that the sentences are truly "pairs", however, this information is probably not difficult to realize by the average participant. For each sentence, the user is asked to agree or disagree (on a classic 5 point Likert scale [Lik32]) with two statements:

1. This sentence is grammatically correct.
2. I feel comfortable using this sentence in my own writing.

6.2.1.3 Pair-wise evaluation of sentences for meaning preservation (survey question 4)

For this question 11 of the 15 pairs of sentences from question 3 were displayed to the user as "A" and "B" options. The user was then asked to assume sentence A (original) is true. Then the user is asked "to what extent is sentence B true?" The answer is given in 4-point scale between "absolutely true" and "not true." The other four sentence pairs not used are interrogatives, and we felt that having a truth assumption about a question may be confusing to the readers.

Question 4 measures the meaning preservation functionality of our system. We would like to replace a sentence with another, which is written differently but expresses basically the same ideas. If this is done correctly, then sentence B would have to be absolutely true based solely on the information that sentence A is true. In other words, they would represent the same content. If the re-written sentence reflects something other than the original, then sentence B is not true or true only to some degree.

A list of all sentences used for Question 3 and Question 4 are presented in the following table. All sentences are used in Question 3, while all but the interrogatives are used in Question 4.

Table 38. All sentences used in survey Questions 3 and 4

Original sentence	Modified sentence	Used in Q4?
Every advance in civilization has been denounced as unnatural while it was recent.	Every advance in society has been denounced as unnatural while it was recent.	Yes
Parents too have expectations and goals which contribute to their own concept of the American Dream.	Parents also have the expectations and objectives that contribute to their own concept of the American dream.	Yes
As far as I know, he's a nice guy.	As far as I know, he's a nice man.	Yes
We have been able to create more effective treatments using this information, as well as preventions and cures.	We have been able to create more effective treatments using this information, also as preventions and cures.	Yes
There is a small chance that he will succeed.	There is a little chance that he will succeed.	Yes
Can I have some anchovies with olives?	Can I have some anchovies with fruit?	No
Do you have something to say?	Do you have a thing to say?	No
The best that I can assume is that all people want a future.	The best that I can assume is that all a future is wanted by people.	Yes
He needs a towel.	A towel is needed by him.	Yes
However, following September 11th, we gave up the right to privacy.	However, after September 11, we waive the right to privacy.	Yes
I use the Internet for business.	I use the Internet for job.	Yes
Hofstader's also points out some of the weaknesses in Turner's thesis.	Hofstader's also points out some of the weaknesses in Turner's dissertation.	Yes
The Great Dane is a breed of domestic dog known for its giant size.	Great Dane is a breed of domestic famous for its giant size.	Yes
Are we having company tonight?	We have company in the evening?	No
Are your parents coming home?	Are your parents coming up home?	No

6.2.2 STYLE SHIFTING (QUESTIONS 5 AND 6)

Questions 5 is designed to test the hypothesis that style changes at the small (sentence) level can lead to changes at the passage or document level. It also tests the theory that style changes as measured statistically by our classifier would be substantial enough to be confirmed by independent human critics according to common

understanding of "style". Question 6 asks for qualitative responses explaining the user's decision in Question 5.

The style-shifting question is the most time consuming part of the survey. For this question, the user is instructed to first read two excerpts from Herman Melville's *Moby Dick*. The user is also given access to a much longer passage from the same book to optionally browse. The excerpts are accessed via HTML links in the question itself. They enable viewing of HTML pages where the text has been prepared. The pages are titled "Excerpt 1", "Excerpt 2" and "Long Text". The author is referred to as "Author A".

Once the texts are read, the user is asked to read two other passages that are much shorter (3 paragraphs each). The question (5) is "which one of the passages is closer to the style of Author A?"

In actuality, the excerpts and the long text are verbatim passages from Project Gutenberg E-book #2711 which is *Moby Dick; or The Whale* by Herman Melville [Mel08]. The E-book is a faithful reproduction of the original classic published in 1851. The *Moby Dick* excerpts and long text were used to train our classifier.

6.2.2.1 Generating passages for style detection

To generate passages, we first need to set up a classifier with some number of appropriate classes. This is so we can detect precisely which class (style) a document belongs to before and after the style transformation operations. We continue this until a source document becomes statistically associated with our model of the target style.

To detect that we can shift toward a particular style, we setup two transformation operations, each shifting toward a different style, and away from the style they were associated with at the beginning. In order to keep as many variables constant as possible, we use the same exact text for both transformations. We also use at least three different classes, so that at the end we can have three texts (1 original and 2 shifted) each associated with a distinct class. The following is a summary of the major documents involved for this question:

Table 39. Documents used in survey Question 5

Document	Description	Shown to user?
AAAC-A1 Corpus	three essays written by the same college student designated as training data for "Author 1" from AAAC corpus, 2004	No
Dewey Corpus	contiguous section from chapters 1-7 of "Democracy and education: an introduction to the philosophy of education" By John Dewey, 1916, digitized as Project Gutenberg E-book #852	No
Melville Corpus	contiguous section from chapters 45 through 58 of "Moby Dick" by Herman Melville, 1852, digitized as Project Gutenberg E-book #2711	Yes, as "Long Text"
Excerpts 1 and 2	2 random, non-overlapping contiguous sections from the Melville Corpus (503 and 828 words respectively)	Yes
Source text	small 2 paragraph section (288 words) from the AAAC corpus "Author 1" test document	No
Text 1	source text style-shifted toward Melville	Yes
Text 2	source text style-shifted toward Dewey	Yes

We represent the shifting from the source text toward "Text 1" and "Text 2" in a two-dimensional setting below. The large, solid circles represent corpora used for training in the classification system. The smaller circles are passages or documents associated with one corpus or another. The broken lines represent a "boundary of

association", that is everything within them is associated with the corpus style or class of the inner circle. In our system, however, we simply define association for a text as having the *closest* distance among all classes in the classification problem.

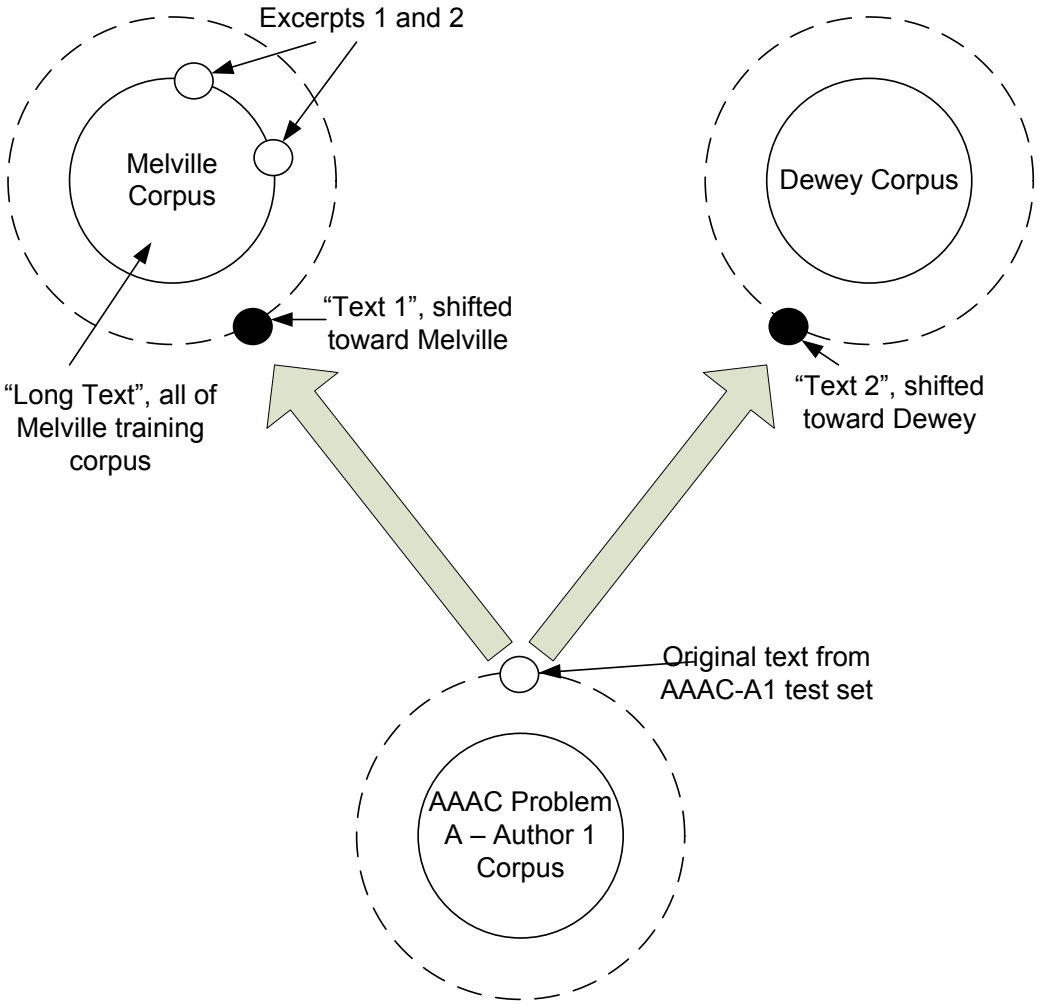

Figure 25. Style shifting in survey Question 5

The dark circles represent the generated texts, which are two different variations of the original. The content for all three variations are displayed the table below.

Table 40. Passage content used for survey Question 5 (Source not visible to user)

Text	Content
Source (original)	Working is something that some people take as a way to accomplish things. Many people take pride in their work and make it their life. Their life revolves around their work. They live, sleep, and eat their work until it is perfect. People with these morals are ideal workers. As years pass we lose more and more people with these morals. People do the bare minimum to get the dollars that they need to survive. Not many people still have pride in what they do, or in the product of their work. In past years quite a long time ago there were people that took pride in everything they did, especially their work. Workers such as carpenters, painters, factory workers, and many other guildsmen, put everything they had into their work. They would put blood, sweat, tears, and most importantly time into their work. It was about pride and respect and they made sure things were perfect in what they produced. People may say that there are still workers of this stature, and there is. Yet they are hard to find. Current day has produced something that no one should be proud of. We have produced money hungry people who focus more on the sale than the actual product. We have produced people who do not take pride in their work, just in that someone compensates them for their work. What makes a great worker these days are the ability to sell not create.
Text 1	Time of the law is that some people take it as a way to get what is passing. Many beings take pride in their work with their lives. And many are doing it. His life revolves around its operations. They be, sleep, and eat their act until it is perfect. People with these property administrations are deadly ideals. As times occur we lose more and more beings with these meanings. Beings are causing, as a lower limit, to receive the buck, which they require to survive. Not many beings have no matter of feeling, what they do, or in a trade good of their work. In the Holocene time periods quite a long time ago, there were people who consumed pride in everything they caused, especially their work. Souls such as carpenters, painters, working persons, and many other guildsmen, put everything they had in their act. They would place blood, sweat, tears, and most importantly passing on into their activity. It was about the sense and element of reference, and things created by them were perfect in the produced. Beings can say, that is, no matter of the workers of this height, and there is. Yet they are

	concentrated by chance. Time produced state of affairs that no one should be proud of. We have created money hungry beings who concentrate to a considerable degree on the cut-rate sale than the actual trade good. We have conveyed forward beings who do not pride oneself in their activity, just in that an individual compensates them for their work. What makes a great mortal, is the opportunity to sell not create.
Text 2	Working instance is a thing that some people take as a state to do things. Their lives rotate around their work. They live sleep and hence their activity. It is perfect. People with these structure powers are ideal workers. Passing class, people are often lost by us with these content. People do the bare peak to get the medium that they need to live. Some people are irrespective of pride in what they do, or the product of their work. In the preceding period of time were people who displayed full pride in everything they did, especially their work. People such as carpenters, painters, factory worker and several guardians, put everything they have in their work. They deployed products, try, tears and most significant instants in their work. It was about pride and respect and asked if everything was perfect in business. People can state that there are still workers and stature. They are still concentrated in order to find. Times made something that no figure should be pleased with. We made medium of exchange hungry people concentrate to a greater extent on selling the actual product. We have prepared the people that not pride oneself in their work, to a man or female person that will cover them for their work. What does a great worker these days, is the power to deal not make.

Lastly, we present the actual distances as measured by our program. Texts 1 and 2 are the final passages displayed to the user for the survey Question 5. The numbers in italics represent the difference between the two texts in relation to the Melville corpus, which is the only one the user is provided writing samples for.

Table 41. Survey Question 5 passages: final distances to all styles

Text	**Distance to AAAC-A1**	**Distance to Dewey**	**Distance to Melville**	**Attribution**
Source (original)	1.336	1.392	1.388	AAAC-A1
Text 1	0.882	0.874	*0.871*	Melville
Text 2	1.697	1.667	*1.685*	Dewey

6.2.3 Detection of computer-transformed text (Questions 7 and 8)

The goal of this question was to assess if computer manipulated texts, specifically ones produced as a result of our style transformation operations, can be distinguished by humans from their original versions.

Question 7 instructs the user to read two passages and explains that one of them is re-written by a computer. The original passage is chosen from the AAAC corpus. The modified version is the result of a style obfuscation operation by our system. Question 8 is a qualitative feedback question that asks the user to state some of the reasons they had for making the decision in Question 7.

Table 42. Passage content for survey Question 7

Text	Content
Text 1 (original)	The irony of the situation is the sympathy our nation has for Monica Lewinski. Her story has portrayed her as somewhat of a victim, while President Clinton is being portrayed as the abuser. Yet she knew very well that he was a married man and that what they were doing could have ruined his career and marriage. She has been a topic on several daily talk shows, and she has made many appearances on late night television. The "exclusive" personal interviews discussing her dramatic situation are countless. She has been the object of Saturday Night Live skits, and the focus on other comical shows. She has even designed her own collection of handbags. To say she has been over-publicized is an extreme understatement. In reality while gaining a negative reputation from the affair she has actually benefited, at least financially.
Text 2 (modified)	The irony of the state is sympathy, our nation sent to Monica. Her story has portrayed her as pretty of a victim, while President Clinton is being portrayed as the abuser. Yet she did it fine that he was a married man and that what they were doing could have ruined his career and marriage. She has been a topic on several daily talk shows, and many appearances have been made by her on late night TV. The exclusive personal interviews discussing her dramatic situation are countless. She has been the object of Saturday Night Live playings, and the heart of other comical shows. She was even designed by her group of handbags. To say she has been over-publicized is an extreme understatement.

	At the same time, wins a negative honor from the things she has learned, at least economically.

6.2.4 STYLE OBFUSCATION QUALITY COMPARISON (QUESTIONS 9 AND 10)

For this question, we are interested in our style obfuscation functionality versus another tool available on the Internet. At least dozens of different tools exist on the Internet, which attempt some variation of this function. Most online tools are advertised as "spinners". The most common use for these is to restate the same English content using different words and sentences for the purpose of appearing to be a completely different text to search engines. But very few of these tools actually meet the criteria that we have in mind. Like ours, we require the tool to be truly autonomous, that is to say, no further human input other than providing the original text should be necessary. The tool should also accept plain text English input and return the same. The tool chosen, and the best that we found, is an online app called "article-rewriter"[Art11] and it is housed in Google servers.

In question 9 we present two articles to the user as "Text 1" and "Text 2". "Text 1" is modified by article-rewriter and "Text 2" is the product of our standard style obfuscation operation, just as question 7. The original is from AAAC corpus (problem B, Author 4). The user was simply asked to choose which text was "written better."

Question 10 is a free form response question allowing the user to discuss some of the reasons behind the decision on question 9. All texts are shown below.

Table 43. Passage content for survey Question 9

Text	Content
Original (not shown to user)	There is no specific date as to the beginning of the Cold War, but its origins begin sometime shortly after World War II. It was after this war that the Soviet Union occupied Eastern Europe, something America was unwilling to accept. America wanted to promote democracy and other American ideas in Eastern Europe. After the Yalta conference, it seemed as though the Soviet Union would be on friendly terms with the Western world, however Stalin was quick to shatter this idea when He refused to reorganize the Polish government in any significant way, suppressed freedom of speech, assembly, religion, and the press in Poland. Stalin followed a similar pattern throughout the rest of Eastern Europe, making it unmistakably clear that the Soviet Union controlled the region, and shut out the West. It was in great part from these actions by Stalin that led President Truman to take a stance on the communist Soviet Union that would later become the main idea of the Cold War, containment. Ambrose defines containment as building up the military strength of America and her allies, and a willingness to stand up to the Russians wherever they applied pressure. America believed that if one country fell to the Soviet Union, the neighboring countries would fall soon as well. This would be the main policy of the United States, and would be upheld by future presidents such as Eisenhower.
Text 1	No date for the start of the Cold War, but its origins begin sometime soon after the Second World War. It was after this war that the Soviet Union occupied Eastern Europe, something the United States was willing to accept. America wanted to promote democracy and other American ideas in Eastern Europe. After the Yalta Conference, it seemed as if the Soviet Union would be a friendly relationship with the West, however, Stalin was about to break this idea when he refused to reorganize the Polish government significantly, suppression of freedom of expression, religion of the Assembly, and the press in Poland. Stalin followed a similar pattern in the rest of Eastern Europe, making it perfectly clear that the Soviet Union controlled the region and set aside the West. It was largely these actions of Stalin that led President Truman to take a position in the communist Soviet Union which later became the main point of contention in the Cold War. Ambrosio is defined as the construction of retaining U.S. military forces and its allies, and the will to confront the Russians wherever pressure is applied. The United States believes that if one country fell to the Soviet Union, neighboring countries would soon be well. This would be the main policy of the United States, and will be confirmed by future presidents like Eisenhower.
Text 2	There is no specific date as to the starting of the Cold War, but its origins begin sometime shortly after World War II. It was after this war that the Soviet Union took Eastern Europe, something America was unwilling to accept. America wants to

	promote democracy and the idea of other Americans in Eastern Europe. After the Yalta Conference, it seemed as if the Soviet Union would be on friendly terms with the Western world, but Stalin was quick to ruin this when he refused to reorganize the Polish Government in any significant way, suppressed freedom of speech, Assembly, religion and the press in Poland. A similar pattern was followed by Stalin throughout the rest of eastern Europe, making it unmistakably clear that the the region was controlled by Soviet Union, and shut out the west. It was in large part from these actions of Stalin, which led President Truman to take a position on the Communist Soviet Union which would later become the main idea of the cold war containment. Ambrose defines containment as building up the military strength of America and her allies, and a willingness to stand up to the Russians wherever they applied pressure. America trade that if one country fell to the Soviet Union, the neighbouring countries would decline rapidly as well. This would be the main policy of the United States, and would be upheld by future President of the United States such as Eisenhower.

6.3 USER STUDY RESULTS

In this section, we present the user study results both quantitative (multiple choice questions) and qualitative (free form responses and interviews). We discuss the results, quantification methods and clarify the choices that were available to the user. We leave the more in-depth discussions about expectations and interpretations to the next chapter.

6.3.1 INDIVIDUAL SENTENCE EVALUATION RESULTS (QUESTION 3)

For question 3, users are asked to agree with 2 separate statements per individual sentence. "This sentence is grammatical" and "I would use this sentence in my own writing." Likert scale response choices are used for both statements. The responses are considered ordinal data and could not be averaged since the relative distinctions between the choices are not continuous to the user. To process them we consider the median of the responses to the original sentence (sentence A), and compare them to the median to the transformed versions (sentence B). We note the users were not specifically given any information about sentence pairing. They were asked to evaluate the 30 randomly ordered sentences as individuals.

The following table summarizes the observations. The vertical axis denotes the (A) sentence ratings on the median while the horizontal axis denotes the (B) sentence ratings. The integers in the cells signify the number of sentence pairs with specific A/B ratings of that location. The number 4 at the cross-section of (A: agree, B: disagree), for

example, means that 4 sentence pairs out of 15 had a median ratings of “agree” for sentence A, and a median rating of “disagree” for sentence B. In other words the transformation caused the grammatical sentence to be perceived as un-grammatical. Four other sentence pairs went from “agree” to “neutral”. Two pairs remained in their same “agree” status while in one case, the original median response was “somewhat agree” and it was actually improved to “agree” after the transformation.

Table 44. A/B median user agreement with "This sentence is grammatical." (15 sentence pairs)

A\B median distributions	B: agree (3)	B: somewhat agree (1)	B: neutral (5)	B: somewhat disagree (2)	B: disagree (4)
A: agree (12)	2	0	4	2	4
A: somewhat agree (3)	1	1	1	0	0

We follow the same process and tabulate the results of the user agreements with the second statement: “I would use this sentence in my own writing.” Here we see a much more dramatic degradation of “naturalness,” with the majority of sentence pairs (9) going from a median of “agree” for A-sentences to a median of “disagree” for B-sentences.

Table 45. A/B median user agreement with "I would use this sentence in my own writing." (15 sentence pairs)

A\B median distributions	B: agree (1)	B: somewhat Agree (2)	B: neutral (1)	B: somewhat disagree (2)	B: disagree (9)
A: agree (11)	1	0	0	1	9
A: somewhat agree (2)	0	1	1	0	0
A: neutral (2)	0	1	0	1	0

6.3.2 *Pair-wise sentence evaluation results (Question 4)*

For this question, the user is presented with both A and B variant of the sentence together and asked to what extent (B) would have to be true if (A) is assumed true. Since this question communicates a continuous variable (degree of truth) and it does not follow a bipolar Likert scale, we analyze the results by calculating the weighted average of the responses. We begin with a weight matrix.

Table 46. Weights of responses for survey Question 4

Response	Weight
(B) is absolutely true	4
(B) is somewhat true	3
There is only a little truth in (B)	2
(B) is not true	1

We obtain the weighted average of all the responses for each sentence pair.

Table 47. Response summary to survey Question 4

Sentence pair	Truth in (B)
1	3.0000
2	2.7895
3	3.7895
4	2.6842
5	3.6842
6	3.0526
7	3.9474
8	3.7368
9	2.6842
10	2.8421
11	2.4737
Average	3.1531

The average degree of truth in B is about 3.15 where a 3.0 indicates "somewhat true" and a 4.0 indicates "absolutely true." We display the results per sentence pair again in the figure below.

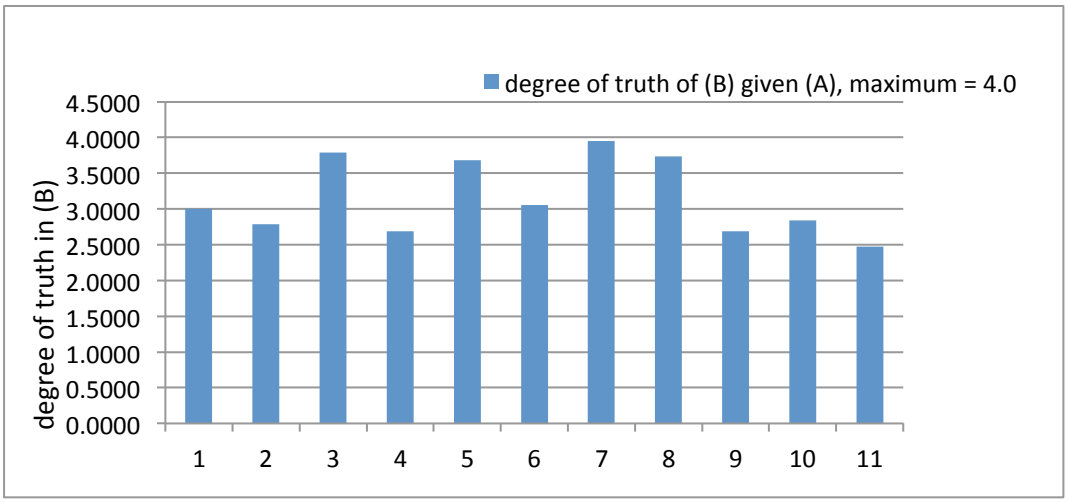

Figure 26. Responses to survey Question 4

6.3.3 Style shifting question results (Questions 5 and 6)

In this question, the users are asked to choose between two paragraphs based on which is closer to the style of texts they were instructed to read. The full details of all documents and generated texts are explained in Section 6.2.2 above.

Out of 19 participants, 12 choose "Text 1" as the style closest to the one they read excerpts of (Melville's *Moby Dick*), and 7 chose "Text 2".

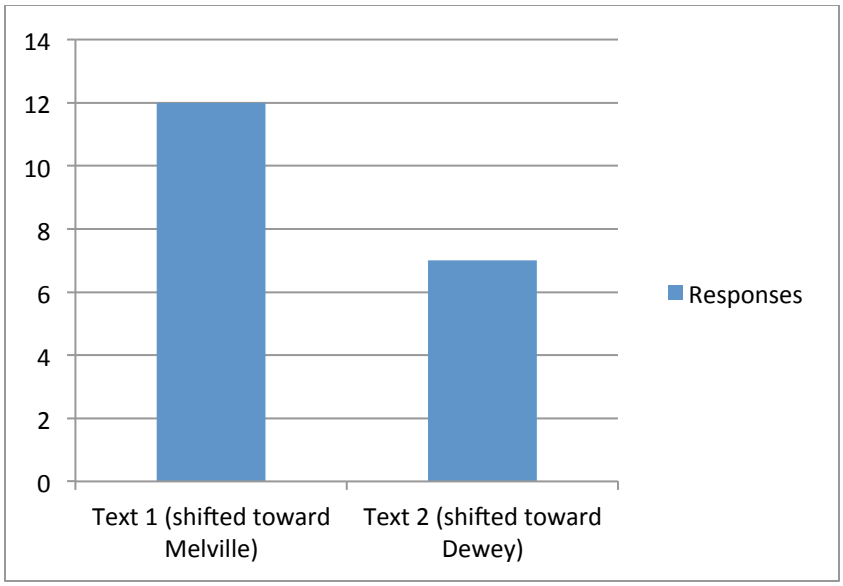

Figure 27. Responses to survey Question 5

For the next question (6), the user is supposed to answer what factors and clues drove the decision in the style shift question. We had hypothesized that factors relating to sentence grammar, "making sense", vocabulary, "natural" sounding and paragraphs may be expected from the users. We tabulate the 19 responses to this question and categorized them in this way:

Table 48. Responses to Question 6 by category

Category	Number of responses
grammar	2
"making sense"	4
words or vocabulary	6
"natural" sounding	3
paragraphs	0
other	6

Of the six responses in the "other" category, "sentence structure" is cited by two participants. Other concepts cited are "fit best", "cryptic" (in reference to the Melville text), "convoluted ideas", "passive voice" and "complex sentences".

6.3.4 Detection of computer-transformed text results (Questions 7 and 8)

Question 7 informs the user that one of the texts was an original, and the other was the computer-modified version of the first. The user is instructed to choose the original. A third, "not sure" option is also allowed. In reality, "Text 1" is the original and "Text 2" is the modified version. The results are as follows.

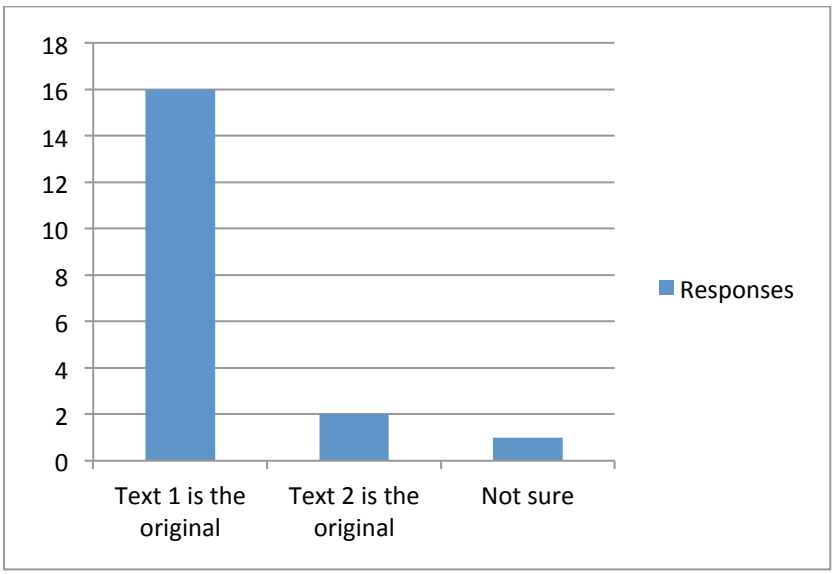

Figure 28. Responses to survey Question 7

Almost everyone is able to correctly guess that "Text 1" is the original. Question 8 asked for the reasons on this decision. We tabulate these based on the same response hypothesis we have for Question 6.

Table 49. Responses to Question 8 by category

Category	Number of responses
grammar	1
"making sense"	10
words or vocabulary	7
"natural" sounding	0
paragraphs	0
other	3

The overwhelming majority of respondents here cited either words/vocabulary or "making sense" as their reasons. Five responses under "making sense" specifically mentioned the phrase "She was even designed by her group of handbags", from the second text as evidence why it was not the original. Other popular reasons included "awkward" or "confusing" words and substitutions. Of the three responses in the "other" category, one cited "personal pronouns", another one states "they're both bad," while the third states "second text sounds like it was written by someone who does not speak English well."

6.3.5 Style obfuscation quality comparison results (Questions 9 and 10)

Question 9 asked the user to simply choose the text that is "written better". Both texts are computer-manipulated versions of the same passage from the AAAC corpus.

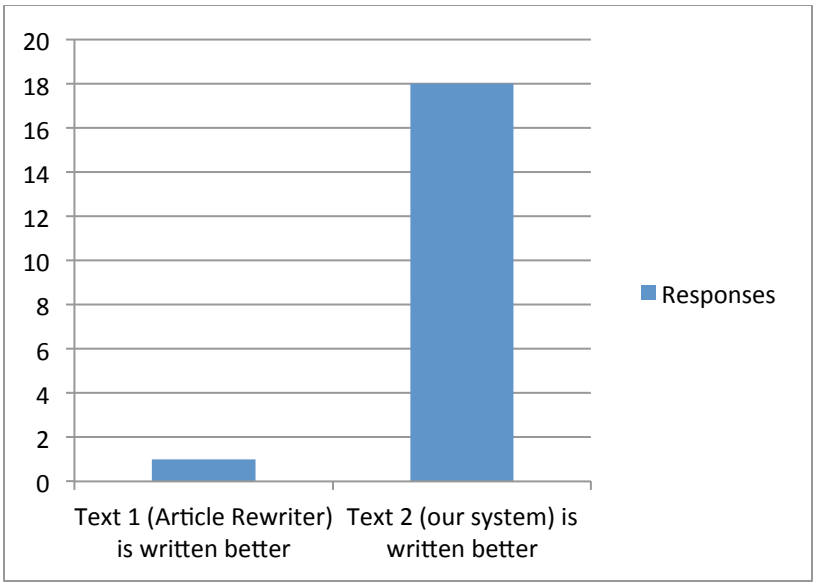

Figure 29. Responses to survey Question 9

Question 10 responses are reasons that users were asked to explain for their decision in the previous question. We categorize these responses using the same classes as Questions 8 and 6.

Table 50. Responses to Question 10 by category

Category	Number of responses
Grammar	4
"making sense"	5
words or vocabulary	2
"natural" sounding	4
Paragraphs	0
Other	1

The responses to this question seem to be almost evenly split between grammar, "making sense" and "natural" sounding. Most people, who cited grammar, also pointed out the grammar problems in the second text, but noted that they were fewer. The responses citing "making sense" pointed to logical inconsistencies between the two

passages and how one is more plausible to be claiming in an essay. For example, regarding Stalin's occupation in Eastern Europe, the first text said that Roosevelt was "unwilling to accept this," while the second text says he was "willing to accept" it. Those who cited the naturalness reason usually referred to the superior "flow" of the second document versus the first. The one response that was tagged as "other" referred to missing words (usually articles) from the first text.

6.3.6 INTERVIEW PORTION

We also conducted short (approximately 5 minutes) phone interviews with 12 of the respondents who replied to the interview request. These provide invaluable insight not captured in other parts of the survey.

In general, our goal is to capture anything that may have been missing in the free form response questions and also to get the impressions of the users who participated in the study. Due to the fact that most interviews took place hours after the written portion, not many users recalled specific reasons for their decisions but gave general impressions instead.

Two questions are asked of everyone in this portion: "What did you think the survey was trying to achieve?" And 2) "What were your overall impressions of the survey?"

During the interviews it became clear that many of the users were confused and puzzled as to the nature of the survey. Some thought there are trick questions and

psychological traps. Many of those who were naturally curious thought the whole point was to distinguish between human and computer generated text. This was confirmed to them once they read Question 7, which specifically states that one of the following texts, is computer generated. However, when told that all the other passage selection choices were between two different computer generated texts, at least four were very surprised. We present a summary of all responses below.

Table 51. Interview responses

User	"What did you think the survey was trying to achieve?"	"What were your overall impressions of the survey?"	Other responses
1	"trying to learn more about different styles"	• "most of the time it was obvious which was the right choice" • "my decision was based on which was written better."	surprised to learn that both choices in Question 5 are computer generated.
2	"no idea"	• "noticed a lot of incorrect English" • "went on which passage has the same voice spanning across multiple sentences."	"choices in question 5 seemed about the same."
3	"to test writing quality"	On the bad passages, it was "not clear what the point of each sentence was"	
4	"study grammar and style"	• "I felt thinking about difference between grammar and style" • "saw one word which I thought was pulled from a thesaurus by a computer."	"had some trouble on style differences"
5	"not sure"	• "Many of the sentences were grammatically correct, but stylistically bad" • "Melville has irregular constructions, so that helped make the decision in question 5"	"experienced a lot of frustration"
6	"my guess: research perceptions of writing styles"		"was very confident in all answers"

7	“don’t’ know”	• first part, the grammar question was “very obvious”, in Question 7, the word “skits” in reference to “Saturday Night Live”, versus “playing” made it obvious. • “Question 5 was very hard,” both passages were "strange" • decision was based on which text was more “connected”	"not very confident on the passage questions"
8	“learn about English styles”	• “I looked for natural and unnatural sentences” • "old" sentence constructions was a factor for Question 5.	
9	“styles”	• "Lewinsky passage had a dead giveaway" • "in hindsight, easy ways to tell human v. computer"	surprised to know that other passages were computer generated
10	“trying to tell bad writing from good”	• “very frustrating” • “decisions mostly based on which one had clearer language.	surprised that both Q5 passages were computer generated
11	“didn’t know until the Monika Lewinsky question, then thought it was all about distinguishing human versus machine writing”	“it was easy to tell which was machine in hindsight”	surprised they were both machines in Q5 and Q9.
12	“reading comprehension”	“some very bad writing”	

6.4 USER STUDY DISCUSSION

In this section, we discuss some challenges encountered during the user study process and discuss their possible effects on the outcome. These challenges are time constraints, subject selection, chance presence of “dead giveaway” sentence constructions and management of user perceptions.

6.4.1 TIME CONSTRAINT ISSUES

By far the biggest challenge in designing this user study was the time constraint. Our algorithms, while sentence based, are evaluated at the document level. The classification exercises all compare documents against collections of documents. For this study, however, we could not possibly expect to find subjects willing to read hundreds of pages of text and making stylistic determinations after hours or days of study. Not only would finding subjects be difficult, but also the lengthy process itself will likely introduce many other uncontrolled factors. Therefore, we decided to use only passage-level texts of only one or two paragraphs for most of the questions. To some extent this was a deviation from our computational process to which we were trying to compare against. To keep the amount of subject's time under sixty minutes, we also exposed the subject to only one corpus (but two unknown documents). The ideal process would have many corpora and many unknown documents to be associated with them. We would then have asked the user to perform style attribution under similar conditions to what our system does already.

6.4.2 SUBJECT SELECTION ISSUES

Participants in this survey were not representative of the population in general. Participants self-selected based on our requirement that they possess "good command of the English language." Most participants were graduate students, with at least five from humanities and literature. This fact by itself is not a problem since this study made no attempt to control for any demographic or educational factors. However, we

hypothesize that educational background, particularly advanced studies involving literature, English or linguistics, was a factor in perception of grammaticality, naturalness and truth preservation. We cannot be sure toward which direction such a bias may have skewed the results.

6.4.3 Handling illogical and unusual sentence constructions

A side effect of the transformation process is that errors are sometimes introduced in the resulting sentence due to unsuccessful word-sense-disambiguation or other factors. These errors are not predictable nor their impact controllable. The addition of a small word may make the sentence logically incomprehensible or change its meaning entirely, while much larger changes may have muted impacts. We call these disproportionally impactful errors "dead giveaways". For example in our user study Question 7, the erroneous sentence "She was even designed by her own collection of handbags", was a dead giveaway. Dead giveaways should of course be minimized in general by having better and more comprehensive paraphrasing methods, however their presence is harmful beyond other mistakes because they make it very difficult to compare the *level* of grammaticality and naturalness of the passage.

6.4.4 User perception issues

Lastly, managing user perception was a challenge. In order not to bias the users, we purposefully did not explain the specific goals of each question to the subjects. But this had the effect of leading the users to speculate about the possible purpose and motivations of the questions. Most participants came to their own conclusions (either

right or wrong) and kept that conclusion throughout the user study. Some of this speculation leads to wrong choices. For example, a few users were convinced that the study is concerned primarily with telling the difference between machine-generated text and the original user composed text. Therefore some "machine-like" mistakes were discovered and used as evidence that the correct response is the other choice.

6.4.5 FUTURE WORK

In conclusion, solving the resource constraint problem around the physical conduct of the user study is paramount to obtaining more accurate and less biased results. Given more resources (in terms of subject time), pairs of transformed texts (such as the one used in Question 5 of the user study) could be doubly verified with access to two separate corpora and two unknown texts. In addition, in the absence of error-free text generation, addressing the "dead giveaway" issue by manually removing illogical sentences could provide a clearer assessment of the overall quality of stylistic transformation.

7.0 EVALUATION

In this chapter, we present our evaluations for the thesis. We conducted many experiments and gathered results in previous chapters. Here, we succinctly revisit the specific hypotheses we presented at the beginning of this thesis and explain the evaluation procedures used to test them, the outcome of those tests, and the significance of the results. In addition, we present evidence of other contributions that were not strictly part of the hypotheses for this work, but were nevertheless instrumental in some aspect of our research.

Section 7.1 of this chapter is organized around the five specific hypotheses that were stated at the beginning of this document (page *xvii*). For each subsection of 7.1, we present the specific tests conducted, results, significance and discussion. In Section 7.2, we provide evidence of other contributions beyond the five stated hypotheses. Finally, in Section 7.3, we discuss and evaluate the overall work of this thesis.

7.1 SPECIFIC HYPOTHESES AND TESTS

We revisit the five main hypotheses in order.

7.1.1 HYPOTHESIS 1: MULTIPLE COMBINED MARKERS WORK BETTER THAN ONE (OR FEW) IN ATTRIBUTION PROBLEMS

7.1.1.1 Tests and results for hypothesis 1

Two specific tests were designed for this hypothesis. The first is:

Test: Compare performance of single-marker systems with hypothetical multiple-marker variants of the same system.

We introduced the JGAAP [Juo09] system in Section 2.8.1. Written to capture the methods used in AAAC event, JGAAP allows the user to choose one "event set" (marker/parameters) to perform classification with. We noted the best JGAAP could accomplish with a single marker was 9 out of 13 correct responses in AAAC problem A (one of the most difficult for the AAAC participants). However, we showed in Section 4.1.1 a weighted plurality algorithm that takes into account predictions from more than one marker could have accomplished a 100% attribution (13 out of 13) on the same problem. We conducted 100 experiments on many possible choices in JGAAP (Appendix C), and were able to select 14 event sets that when properly weighted showed the perfect result.

The second test for this hypothesis was:

Test: Generate own marker organization allowing for combinations. Use it to build multi-marker attribution/classification system. Show broad attribution performance gets better with more markers.

We presented a full taxonomic system as our marker organization in Chapter 3. Later in Section 4.1.2, we used a first-choice Hill Climbing greedy method together with a Euclidean distance formula to combine multiple markers in a weighted fashion (a more sophisticated algorithm for accomplishing a similar task that we did manually in

4.1.1 with JGAAP results). We experimented with style attribution using the Brown categorized corpus. 12 multi-marker attribution experiments were conducted. We reported in Figure 6 that successive experiments with increasing numbers of markers lead to better classifications. The performance jumped from an average of 16 correct attributions (14 markers) to 22 correct attributions (232 markers).

7.1.1.2 Significance of results for hypothesis 1

The first test was a hypothetical toy experiment that was conducted manually. Since all the attributions were known ahead of time, the test updated a theoretical upper bound of performance. It only demonstrated that an authorship attribution system, which was designed around using just one event set at a time, could be significantly improved with the use of a multiple-marker ensemble method. As such it verified the hypothesis in an unbiased manner.

The second test was a true attribution problem using the first-choice Hill Climbing algorithm to train the classifier and a k-NN Euclidean algorithm for distance. It may seem somewhat intuitive that more markers improve the machine learning system, but the converse has been reported in [DeV01] and [Li06]. In [DeV01], the attribution system suffered with the number of markers increased from 122 to 320. Authors in [Li06] devised a sophisticated Genetic Algorithm (GA) based feature selection phase that would vary the set of features used in their ML algorithm to find an optimal set. They found that an "optimal" feature set improved the accuracy of their SVM-based classification compared to the "full" feature set: from 97.85% to 99.1%.

However, we have not observed any degradation in performance as a result of increased number of features, even though we clearly have many similar and non-contributive features which [Li06] cited as a reason to use feature selection criteria. Our argument is that if an optimal weight distribution algorithm is used, then by definition, the non-performing markers will be weighted zero or arbitrarily small and therefore would not cause a reduction in accuracy. In other words, there *already* is a "feature selection" aspect built into any multi-feature learning algorithm and there should be no need for a separate GA phase for this. For example in our first test, we manually assigned weights for the weighted plurality algorithm. The vast majority of event sets were weighted zero since they did not contribute to a better-combined result. It is not possible for more event sets to have degraded our approach. We fully understand that implementations may not always be optimal or resource constraints may be a bottleneck as [Li06] also discussed.

While we observed improving accuracy with larger marker sets, the marginal return for each additional marker was clearly decreasing. It is logical to assume that once an optimal solution is reached, then additional markers cannot contribute to more accuracy. But for this to occur, we would have to have already reached 100% accuracy, which we did not in our second test. We cannot ignore any feature before 100% accuracy is reached.

Thus, our results provide additional evidence toward the theory that more markers lead to more accuracy.

7.1.2 HYPOTHESIS 2: SOME MARKERS ARE MORE "UNIVERSAL" THAN OTHERS. THEY WILL STAND OUT IN WIDER ARRAY OF PROBLEMS AND THEREFORE BE CONSISTENTLY MORE INFLUENTIAL THAN OTHERS.

7.1.2.1 Tests and results for hypothesis 2

The test we devised for this hypothesis was:

> *Test: Show experimentally that as the number of corpora/styles grows, some markers are relatively persistent (i.e. consistently highly weighted) across attribution problems.*

We covered an extensive individual evaluation of over 500 markers in Section 3.6 and found that certain markers can act as discriminators in a wide variety of classification problems. In Section 4.3.1 (effects on markers with increasing number of classes) we showed how a number of "contributive" markers are highly weighted by our system across multiple experiments. In successive attribution experiments with higher and higher numbers of classes, we found a small group of markers that are more persistent than others. Table 4 lists six markers that were contributive to 7 or 8 different experiments each with different class-sets.

7.1.2.2 Significance of results for hypothesis 2

We show the significance of these results by comparing them to a chance baseline obtained by simulation. Table 6 lists the observed number of contributive markers in experiments with increasing number of classes. There were 4 trials for each class-set resulting in 48 total trials. To obtain the baseline, we conduct the same

experiments again, except that for each class-set, we replace each document's marker values with a random permutation of the originals. The permutations pseudo-randomize the values of each marker, but keep the overall distribution unchanged. We conduct this version of the experiments five times, for a total of 240 trials.

From these results, we derive five sets of frequency distributions giving us a mean (μ) and a standard deviation (σ) for each frequency value. In Table 53, x is the frequency of contributive markers in different class-set experiments. Specifically x stands for *highest number of classification experiments with distinct class-set for which a marker was a contributive marker in at least one trial*. For example at (x=2), 25 markers were observed, meaning 25 markers were contributive in exactly two groups of experiments, where each group has a unique number of classes.

Table 53 also contains the baseline (pseudo-randomized) average frequency from the 240 trials, along with the standard deviation per x. The right-most column is the calculated absolute value of the Standard Score (z-score) that determines significance between the observed and baseline numbers of markers per frequency. For (x=2), for example, the z-score is 2.33, meaning the observed values were 2.33 standard deviations away from the baseline average.

The z-score values for ($x \geq 5$), are undefined. This is because no marker fell into that frequency during our 240 pseudo-randomized trials and therefore no baseline standard deviation was available. Had we conducted exponentially more trials, we may have seen small numbers appear in this range that would have allowed us to compare

against observed numbers of markers. However statistically, it is clear that for higher values of x, the average μ is approaching zero, and z-scores are becoming larger as markers of x frequency become rarer and rarer.

Table 52. Observations and baseline for hypothesis 2

x	observed number of markers (m)	baseline average number of markers (μ)	baseline standard deviation (σ)	absolute z-score $= \left\lvert\frac{m-\mu}{\sigma}\right\rvert$
0	385	303.80	7.19	11.29
1	88	175.40	6.53	13.38
2	25	42.80	7.63	2.33
3	11	7.80	0.75	4.27
4	7	0.20	0.40	17.00
5	4	0.0	-	*undefined (no σ) but significant*
6	4	0.0	-	
7	3	0.0	-	
8	3	0.0	-	
9	0	0.0	-	*Undefined*
10	0	0.0	-	*Undefined*

For ($x \geq 3$), the z-scores are above 4.0. If we assume a Normal distribution, this z value is already below 0.006% (4σ). Thus, the 32 observed contributive markers that have x values 3, 4, 5, 6, 7 or 8 would be significant at a $p < 0.00006$ level.

In testing hypothesis 2, we have shown strong evidence that there are some markers that remain relatively highly weighted across different problem sets, and that this number is significantly above chance as shown in our comparisons to a pseudo-random baseline.

However, much more could be done to strengthen this particular hypothesis. For example we chose AAAC problems A and B because their documents were made

more uniform by the virtue of their purpose. The essays were from a pool of students, generally about the same subject and written in the same time period. These and other features of the problem greatly help researchers keep other variables constant while experimenting with one (in this case number of classes.) Also, these problems had the highest number of classes (authors) out of all the AAAC problems.

Further, we could compare the number of markers themselves to the subset that we call "contributive." We had 530 in our pool of markers. Would having 300 or 200 markers necessarily mean less number of contributive markers proportionally? Or would we, for example, see more markers become significant in a less crowded field? These would have made for more interesting experiments that could have been done.

7.1.3 Hypothesis 3: Machine classification based on our statistical model of style correlates with human classification based on style (common understanding).

7.1.3.1 Tests and results for hypothesis 3

The test for this hypothesis was:

> *Test: Human evaluators to classify some documents, given the same corpora and classes, and show correlation with machine classifiers.*

In a general sense, to the extent that we have a good classifier, it goes to show that humans agree with the classifications. That is how a "good" classifier is determined: On its level of agreement with human labeled corpora.

More specifically about style-based classifications, where the humans are specifically instructed to choose based on "style", then a much more controlled experiment is needed. We designed that experiment as the user study Question 5 (Section 6.2.2).

The subjects were given the entire training corpus for *Moby Dick* and were presented with two style-shifted choices generated by our transformation algorithms. One of the passages was classified as *Moby Dick* style by our classifier while the other one was not. The results as reported in Section 6.3.3, show users correctly picked the right passage by a 12 to 7 margin.

7.1.3.2 Significance of results for hypothesis 3

Participants were asked to choose between two passages for this problem, one of which was closer to *Moby Dick*, according to our model. A random baseline for this problem would be 9.5 out of 19, which is lower than our 12 out of 19. According to the Binomial distribution ($p = 0.50$ and $n = 19$), $f(12; 19, 0.5) \sim 0.0961$, which means only about 9.61% of the time, this outcome would happen by chance, given a random guess on the part of every user. Cumulative probabilities of all outcomes 12 to 19 is about 0.1796, meaning 17.96% of the time, we can expect a chance outcome to be 12 or higher. The standard deviation on the $B(19,0.5)$ distribution is 2.179. This places our observation between one and two standard deviations away from the mean.

However, it is still difficult to assess how significant this finding truly is. Several reasons for this are perhaps obvious.

1. The human notion of style is very subjective, and the chances of 3 or 4 individuals having conceptions of style outside of what we have modeled are probably significant.
2. There are very few studies in the literature that have user-evaluations of multi-paragraph length text for non-functional language aspects. It's difficult to know what the research community in this field considers an acceptable standard.
3. It is difficult to create passages different enough to exhibit unique stylistic differences from styles that are relatively close to each other. But styles that are far apart from each other, also introduce other factors that could bias the results.
4. This is one trial with one pair of passages, whether this extrapolates to all or many typical passages is highly unclear.

We take each of these reasons and assess them in turn before making recommendations for improvements in the future. We use a three point scale as "likely a contributing factor", "neutral" and "not likely a contributing factor" for each of the objections.

Reason number (1) is correct. Users do not necessarily have the same notion of style as we do. This was proven by their own responses in the interview and free-form questions.

After reading the responses to the free-form question (Question 6) and the discussion during the interviews, it became clear that a significant number of users actually used a different criterion to select between the two passages. This criterion can perhaps best be labeled "clarity." Table 49 lists 7 users that indicated "making sense" or "naturalness" as a reason for their decision in Question 5. Furthermore, in user responses 1, 7, 8, 10 and possibly 3 outlined in Table 52, some version of "correctness" or "clarity" was provided by the users. In our model of style, we do not directly measure clarity and we have no specific marker for it, although several could be argued to be related. We know from our own transformation processes that there is significant chance for mistakes in the generation process and indeed both passages had grammatical mistakes. However, the correct passage did sound more awkward according to the feedback from the people for whom this was an important criterion. This is partly due to the fact that Melville himself, writing a work of fiction in Victorian prose reads less clear than Dewey's work of non-fiction written over 50 years later. Therefore several pieces of evidence, including user testimonials, point to users choosing the wrong passage because it sounds "clearer." Thus we can say with high degree of confidence that reason (1) was a factor for some individuals who chose the Dewey-style passage, but less likely to have been a factor for people who chose the (correct) Melville-style passage.

Reason (2) is also demonstrably true. Our own survey of related literature revealed very few sources of passage-level monolingual user studies, where an artificially induced non-functional aspect of writing, such as style, tone, mood or voice

was being tested. One such study however was very instructive. The study by Inkpen et al., "Generating more –positive and more-negative text" [Ink06] sets up an experiment similar to our user study Question 5 in many respects. The authors selected texts from the British National Corpus (BNC) and an online opinion feedback site (e-pinions). For each text they generated a "more-positive" and a "more-negative" version of the same passage. The techniques and databases used are interesting and we discuss them in some detail in the background Section 2.9.1 ("Related work in concept-to-text paraphrase generation"), but here we concentrate on the user study part of it.

We used six judges to evaluate six *original/modified* pairs of passages and determine whether they were more positive, less positive or equal in semantic orientation. Each pair of texts was evaluated by three judges. Averaging each of the evaluation sets, we found that in two cases out of six, the net-evaluation was correct, i.e. in line with the intended semantic orientation. In three out of six cases, the evaluation indicated the opposite of the intended effect. Finally, in one case, the evaluation was neutral, with two opposite-voting judges canceling each other and the third voting neutral. The results were not conclusive as admitted by the authors in [Ink06]. In the discussion, they cited three main reasons: the scale of the experiment was too small; the numbers of near-synonyms (potential points of paraphrasing) in their chosen paragraphs were too few; and the instruction to the judges was too brief.

In two out of three cases, we avoid the issues stated in [Ink06] and in one case, we arguably repeat them. Although, we conducted the user test with 19 users (judges),

we have not gone very far beyond [Ink06] in terms of scale. We can roughly calculate scale by the total number of "judgments" made by different people on different texts. We multiply the number of judges, J, by the number of unique experiments, E, to obtain a formula:

$$Scale = JE$$

By this calculation [Ink06] had a scale of six judges multiplied by three experiments = 18 judgments, while we had 19 judges multiplied by one experiment for a total of 19 unique judgments. However, in our case, there was about thirty minutes of preparation required to answer this question alone, which made it much more difficult to involve more people or more texts. Nevertheless, this is the pitfall we fall into as well.

We avoid the other two issues raised in [Ink06]. We agree that at least judging by the sample provided in the paper, the number of replaced words appears to be far too small to produce the intended effect. We calculate that in a passage of 330 words, [Ink06] only replace 10 words with heir "more positive" or "more negative" near-synonym. By contrast, every single sentence in our two passages for Question 5 has gone through at least one change and many have seen 6 or 7 different changes. Lastly, we provide more instruction to the user and possibly clearer instructions. In addition to explicit instructions, the user is more familiarized with what we're asking just by reading the three extra passages from Herman Melville.

In total, on reason number (2) we offered above, we say that this issue could have affected the outcome of this study. If possible, we should have many more judges and multiple distinct texts for judgments in the future.

The next reason, (3), was about the difficulty of generating good passages that allow reasonable evaluations while at the same time exhibiting noticeably different styles. We actually encountered this problem early on in our own research. At first, we had decided to use the AAAC corpus not just for the source text but also for the target style profile that we transform towards. Our first attempts to design Question 5, was using AAAC problem A "authors" instead of Melville and Dewey. We wanted to present the users with the work of a AAAC author and have them assess which passage would be closest to it. Because the AAAC problem A corpora are only short essays (3 training essays per author), it was difficult to generate a style very distinct from the source text itself. We received feedback that the two passages are not very distinctive enough and choosing one of them would be arbitrary. In order for our transforms to produce replacement sentences of more variety, we had to develop much larger corpora before attempting the style shift toward one of them. Thus we arrive at the present configuration of the user study with works of Dewey and Melville forming the two distinct styles that we transformed toward.

Another shortcoming with the generation problem is the overall issue of control in text generation. By its nature, there is much less control over the expressions in a text-to-text generation paradigm, compared to concept-to-text, where much more

freedom is available. We can point to the possible inadequacy of our seven style transforms for this issue as well. Ideally there should be hundreds of transforms, allowing for many more possibilities in reformulating the text and therefore providing greater stylistic control over the final passage.

Our last reason (4) is valid as well. Only one pair of texts were used with this user study, mainly due to resource restrictions and time commitment required for this kind of evaluation. Future work should explore this problem and find creative solutions to incorporate a larger sample of texts for user evaluation.

Table 53. Summary of hypothesis 3 evaluation discussion

Issue	Evidence	Assessment
Conception of style may be different than ours	1. User feedback indicated a significant minority chose based on "making sense" or "clarity", rather than style 2. Most users who acted on "clarity" chose the wrong passage	likely a factor
Few studies in the field lack clear standard	Inkpen's study provided some reasons for their inconclusive results [Ink06] 1. Scale was too small (6 judges x 3 passages): By this standard, our study is also limited, but we're not clear where the threshold is. 2. Not enough changes to text: Our study had many more changes compared to [Ink06]. 3. Judges not instructed enough: We believe we had much more instruction both from written directions and reading passages.	Based on very limited field research, the scale (1) was likely a factor in our results. Reasons (2) and (3) were not likely factors.
Difficult to generate good passages	1. Our own early assessment indicated the same problem, thus we switched to a more expressive set of corpora. 2. Overall problem still remains (i.e. with "clarity" becoming an alternative criterion) 3. Limited number of style transforms limits the total control over generated text.	We improved over initial state and in comparison with [Ink06], thus (1) and (2) were not likely factors in the final product.

		(3) is likely a factor.
Only one pair of passages, hard to generalize	only one pair of passages used due to resource restrictions	likely a factor

7.1.4 HYPOTHESIS 4: STYLE SHIFTING AT THE SMALL UNIT LEVEL LEADS TO STYLE CHANGE AT THE LARGER, DOCUMENT LEVEL.

This hypothesis makes a complex argument. It says that unit level style-shifting leads to document or passage level style shifting. We have already demonstrated style shifting statistically according to our model: First with two toy examples in Section 1.6, and second using the Phrase transform and JGAAP in Section 5.2.5.4 and 5.2.5.5. As we argued in Section 5.2.5.6, what has been missing is proof that the unit level style shifting had taken place. While we were in general aiming to produce correct language, reflective in meaning of the original source, this part was never verified formally.

With this hypothesis, we put additional constraints on the transformation system to produce unit-level style shifting. According to our model, style shifting occurs when different, equally as meaningful options are substituted for an original lexical-syntactic choice and the result retains the same meaning. Two constraints, specifically designed to ensure this condition are a) grammaticality of the resulting transformed sentence and b) meaning preservation of the transformed sentence.

7.1.4.1 Tests and results for hypothesis 4

Here the test was:

> *Test: Create a system to strategically apply style transforms given a library of them. Use human evaluators to verify that system unit-level transforms are a) grammatically correct and b) semantics preserving. If (a) and (b) are met, then we have accomplished unit-level style shifting.*

We created the transformation system with a library of seven transforms (Chapter 5) and an algorithm for their successive application to text. This transformation system can strategically choose sentences for replacement that further its style shifting aims. This system itself is unaware of the quality of sentences; it only operates to maximize the statistically driven style goals. Thus to take advantage of unit level style shifting, we must make sure that the sentences available to the transformation system (i.e. from individual transforms) meet the (a) and (b) criteria above.

The transforms were applied to a collection of random sentences and the results were evaluated by humans in user study Question 3 and Question 4. In Question 3, the results showed that most sentences modified by our transformation system degraded with respect to grammar. In four cases the degraded sentence were deemed ungrammatical while in another four cases they were deemed neutral by the users.

In user study Question 4, the users were asked to rate whether sentences retained their meaning after the transformation. The average post-transformation sentence was judged to have 3.15 degrees of truth out of a possible 4.0. The original sentences were pegged at 4.0 (“absolutely true”) for this evaluation. We can conclude

that truth is preserved -not absolutely- but to a moderate degree (3.0 level was labeled "somewhat true" on the user study) as a result of our stylistic transformations.

7.1.4.2 Significance of results for hypothesis 4

The results in the previous Section showed that unit level style shifting operations degrade both grammaticality and meaning. These results are based on significant number of trials. For Question 3, 570 individual human assessments were made (30 sentences x 19 users). For Question 4, 209 assessments on sentence pairs were made (11 pairs of sentences x 19 users). The issue is to what extent are the degradations acceptable? That is: How grammatical does a transformed sentence need to be considered "grammatical"?

We argue that two baselines should be met. First for both the grammatical and the meaning-preservation questions, the random choice baseline must be exceeded. This is a minimum baseline that verifies user responses are above noise. Second, the level of score assigned for grammaticality should be above "neutral", meaning we should show that a majority of transformed sentences, even if degraded, are still judged to be at a neutral or higher level of grammaticality. For the meaning-preservation question, it should also be above 3.0 average (but on a 4 point scale), or "somewhat true" as stated in Table 47. Note that the random baseline for the grammaticality question is also at 3.0 (average of choices 1 through 5), but for the meaning-preservation question, it is at 2.5.

We first examine the response patterns to Question 3. The following table lists the total responses given by all users by Likert scale categories.

Table 54. Distribution of responses to Question 3

	agree	somewhat agree	neutral	somewhat disagree	disagree
A sentences	203	25	13	12	32
B sentences	82	33	34	37	99
all sentences	297	60	48	51	114

We observe that the pattern is far from random as the null hypothesis would indicate an even distribution of responses for all sentences. Instead we find that the distribution for all sentences and A-sentences are heavily skewed toward "agree" end of the spectrum while the B-sentence distribution is more bipolar.

To assess whether or not the degradation in grammar is acceptable, we examine Table 44 and summarize its finding further in the following table.

Table 55. Summary distribution of B sentence median responses to Question 3

	agree or somewhat agree	neutral	disagree or somewhat disagree
B sentence medians	4	5	6

These results indicate that slightly more B-sentences (post-transformation) were judged to be wholly or partially ungrammatical versus those judged to be grammatical or somewhat grammatical. Thus, strictly speaking, we fail to meet this

particular baseline for Question 3 results. That is, the degradation of grammar as a result of our sentences is higher than our expectations.

The second part of this hypothesis was about preservation of the meaning in sentences. We designed the user study Question 4 to assess this. On a 4 point scale, where 4.0 stands for “absolutely true” and 3.0 stands for “somewhat true”, we asked the users to assume statement A is true, and assess the truth of statement B. If both sentences held the same meaning, then B would be absolutely true by the virtue of A being true. But if there is a deviation from the meaning, the truth of B may degrade or not be decipherable from A alone. The average rating of B (3.15) is 5% above our designated baseline, and 26% above a random choice baseline (2.50).

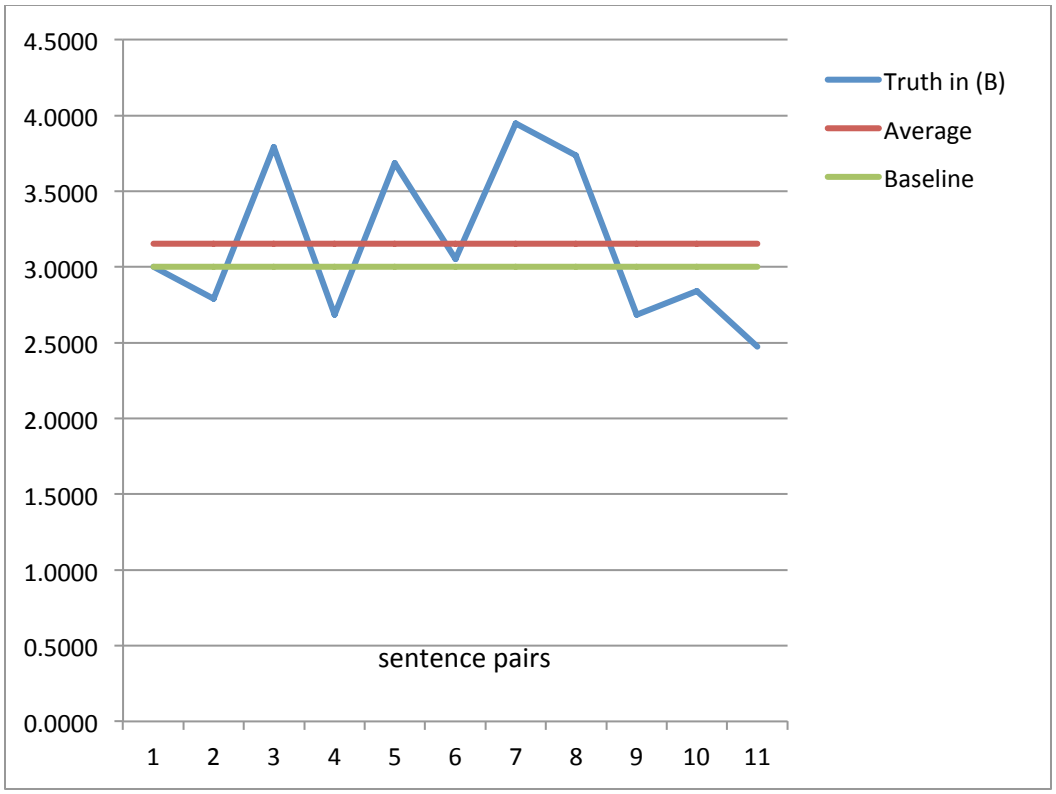

Figure 30. Graph of user responses, average and baseline for Question 4

We have shown in this section that we have met the grammatical and the meaning-preserving criteria for showing unit-level transformation. This realization in addition to the statistics based style transformations that were already done, shows that unit-level style transformations can lead to document level transformations. Further direct evidence is provided by the user study Question 5, where two transformed versions of the same passage have been classified as different styles.

We have shown average performances above the random choice baselines. We assess the grammaticality performance to be low and below our own baseline and the

meaning preservation performance to be moderate at best. We assess that we are providing only moderate evidence for the validity of this hypothesis, given the current state of the technology.

As we see in Figure 22, the most precise transforms are also the ones with the lowest recall. This means any transformed sentence we encounter is many times more likely to have been the product of a lower-precision transform, rather than higher ones. A richer library of transforms, and ones with more inherent accuracy would be necessary to improve the results.

It is also interesting to note that the lower performance on grammaticality did not translate to as low a performance in meaning preservation. This could indicate that correct grammar is not that important to style or at least less important compared to meaning preservation. This hypothesis also should be tested in the future.

7.1.5 Hypothesis 5: Incremental transformation from one style to another is possible

This hypothesis is in a sense the culmination of the entire thesis. And almost all of it has already been demonstrated in previous hypotheses and experiments.

7.1.5.1 Tests and results for hypothesis 5

The test for this hypothesis is:

> *Test: Create libraries containing a variety of transforms and markers from original contributions and several different sources in the literature. Use the same*

style classification method in a typical classification exercise, and show that it will re-classify a document after a series of transforms have been applied to it. Demonstrate both targeted and anti-targeted varieties. Successful machine transformations to be verified by humans, given the same corpora and classes, in a user study.

The libraries containing multiple markers (530) and multiple transforms (7) were enumerated in Chapters 3 and Chapter 5 respectively. A number of transforms have been done with either the full system or partial demonstrations. We exclude sentence-level transformations used in Chapter 5 to verify individual transforms, and list the most instructive document/passage level transformation below.

Table 56. Summary of major transformation experiments

Transformation	Reference	Markers and Transforms
Two small toy problems for demonstration	Section 1.6, "modernizing Shakespeare" and "A trivial 2-marker scenario"	two markers and two transforms
Style obfuscation with AAAC problem A evaluated by JGAAP	Section 5.2.5.4, and [Kho10-2]	one set of markers (JGAAP *Words* event set) and one transform (Phrase)
Targeted style transformation with AAAC problem A, evaluated by JGAAP	Section 5.2.5.5 and [Kho10-2]	one set of markers (JGAAP *Words* event set) and one transform (Phrase)
AAAC passage transformed to styles of Melville and Dewey for user study	Section 6.2.2, user study Question 5	530 markers 7 transforms
AAAC problem B, Author2 passage transformed to "Author1"	Section 6.2.3, user study Question 7	530 markers 7 transforms
AAAC problem B, Author 4 passage obfuscated for comparison with *article-rewriter* tool	Section 6.2.4, user study Question 9	530 markers 7 transforms

The process for generation of the passages for user study Question 5 (Section 6.2.2) accomplishes the statistical transformation part of the question. One passage from AAAC that logically belongs to its own class was transformed in two different ways resulting in two passages that are placed in two separate classes. Furthermore, the result was verified by 12 of the 19 participants of the user study in that they agreed one of the passages (previously in the same class as the other) was now in the *Moby Dick* style class.

To address the *incremental* part of the hypothesis, we present Appendix D, which is a high level log file of a transformation operation. The system runs through every sentence of the initial source passage. For each sentence, there are one or more

transformed replacements suggested by the transformation system. If the new sentence closes the distance to the target style, it is "adopted" and replaces the old sentence. Otherwise, the system moves on to the next sentence. The distance to target is recalculated after every replacement, and a full classification of the text is done after every paragraph. If the classification indicates the desired target class, the system declares transformation success. If a full cycle passes (meaning every sentence has been examined once) without any replacements, the system concludes that no more changes are possible and halts.

7.1.5.2 Significance of results for hypothesis 5

The human evaluation results for hypothesis 5 are identical to the ones for hypothesis 3. They both analyze the results of user study Question 5. We refer the reader to Section 7.1.3.2 for a detailed discussion about the results and their significance.

7.1.6 SUMMARY AND SELF-ASSESSMENT OF THE MAJOR HYPOTHESES

We present the following summary table and self-assessment from the five specifically stated hypotheses. The self-assessment assigns a confidence interval within 0.0-1.0 to each hypothesis. The closer to 1.0, the more confident we are in the evidence we have shown for the validity of the hypothesis.

Table 57. Summary and self assessment of major hypotheses

Hypothesis	Demonstration	Self-Assessment
1. Multiple combined markers work better…	We showed how JGAAP improves with multi-marker variant. Experiment on Brown corpus showed attribution improving with more markers	fairly conclusive, but some contrary studies exist **[0.7-0.9]**
2. Some markers are more "universal" than others…	Showed certain markers to be contributive to multiple AAAC attribution problems at a significant level above a chance baseline.	fairly conclusive **[0.7-0.9]**
3. Machine style classification correlates with human…	Results from user study Question 5 is positive and significant at $p = 0.18$.	moderately conclusive given limited scale (see discussion in 7.1.3.2) **[0.6-0.7]**
4. Small, unit-level style shift leads to overall style shift…	Sentence-level style shift was shown with 570 human assessments on grammaticality, and 209 human assessments on meaning-preservation. Results were above random baselines, but below the baseline for the grammaticality question. The results were above the baseline for the meaning question. Same sentence-level tools were used in successful passage-level transformations.	grammaticality results are considered low, but still above baseline, overall confidence: **[0.35-0.55]** * The importance of grammar to style processing requires further assessment.
5. Incremental style transformation is possible…	Incremental nature of transformation demonstrated in Appendix D. Human verification of style transformation from Question 5 is 26% above random chance average.	fairly conclusive on the incremental question, but limited scale of study (see discussion in 7.1.3.2) **[0.6-0.8]**

7.2 OTHER CONTRIBUTIONS

In this section review other contributions outside of the five specific hypotheses that we presented at the beginning of this thesis, and evaluated in Section 7.1.

7.2.1 EXTENSIBLE, MODULAR, HETEROGENEOUS STYLE PROCESSING SYSTEM

We have built a prototype of the system described in Section 1.3 and Figures 1 and 2. At the heart of this system is the dynamic interaction between a statistical classifier and a multi-layered paraphrase generator. Our survey of literature has

revealed numerous text classification systems used in research, many of them in very mature stages. But most of them are designed to either present useful information (about classification) to the human operators or accomplish classification-centric tasks like spam control and plagiarism detectors. None are exclusively designed to detect incremental style differences. At the same time, there are many language generation and paraphrasing systems in the literature but not many with dynamic goals automatically extracted from the text, and none with a particular outcome of a classification system as a goal.

Table 58. References in two main areas of literature

Area	References
Text classification	[Abb05], [Abb06], [Abb08], [Arg98], [Arg03], [Baa96], [Bre09], [DeV01], [Fak01], [Hol05], [Iqb08], [Jia97], [Juo04], [Juo06], [Juo09], [Juo10], [Kes03], [Li06], [Luy04], [McE00], [Mos84], [Rao00], [Ste09], [Whi04], [Zhe06]
Natural Language Generation and paraphrasing	[Art11], [Bar01], [Bri00], [Bou11], [Cos11], [Dim94], [Ger00], [Gon07], [Hov88], [Ink03], [Ink06], [Jac97], [Kac06], [Kes10], [Lan94], [Lan02], [Lav97], [Loe96], [Mai08], [Mai10], [Sid10], [Sci05], [Wal02], [Wub11], [Zha07], [Zha09], [Zha10], [Zhu10]

Our prototype, used to prove hypotheses in this thesis, brings together the two worlds of classification and generation and makes them interdependent in style processing operations.

In addition, we have designed this prototype to have the following features:

- Extensible: Additional functionality can be added without problems or invalidating previous modules.

- Modular: The system is decentralized and independent of the number and type of modules currently available to it.
- Heterogeneous: Different theories and assumptions can co-exist inside the modules that handle distance calculations, marker extractions and transforms. The system can handle overlapping functionality and redundancy. Unlike most generation and classification systems, the individual decisions (for example around pre-processing of text) are pushed down to the module level, such that they need not be uniform throughout the system.

We believe this system to be a unique contribution. Not only for the approach it takes to do processing, but also physically as software that achieves style transformation by example, validated by text classification.

7.2.2 STYLE MARKER TAXONOMY

The style taxonomy presented in Sections 3.4 and 3.5 is not a unique concept in and of itself. Many sources have provided similar taxonomic categorization implicitly or explicitly. As such we are offering another classification system with different goals and assumptions. However, we believe we have one of the more comprehensive taxonomy of markers, organized around implementation methods. One addition we made was to include common processes such as text preprocessing and outcome summarization directly as separate marker instance categories making them totally unique and therefore, as we argue, comparable and classifiable in discussions.

Hitherto, style markers have been considered machine learning features and discussed in the same sense with the same vernacular. We believe this taxonomy is a more useful way of thinking and exchanging ideas about style markers.

7.2.3 Evaluation of over 500 markers

In Section 3.6, [Kho11] and Appendix B, we presented the results of our experiment with 502 markers evaluating 187 documents of diverse type and authorship in 8 English-language AAAC problems A, B, C, D, E, F, G and H.

We evaluated the results by three criteria related to a style marker's ability to act as an effective discriminator in multi-class classification problems of the AAAC. These criteria are absolute performance (in terms of the number of classified documents), relative performance and comparative performance by average z-scores.

The results are in Section 3.6.2.

7.2.4 Evaluating Microsoft Translation languages for paraphrasing

Using automatic translation tools to generated paraphrases through so-called pivot translations (translating to a middle language before the final one), has been done many times in the literature and is not novel. However, in designing our Translation transform (5.2.7), we used human judges to validate which of the Microsoft Translator supported languages are best suited for paraphrasing by English to English pivot translation.

In addition to conducting the study, we present a novel scoring scheme for paraphrase-by-translation systems and introduce a "goodness ratio" by which we ultimately pick a chain of languages to use for our translation transform.

The study was conducted with two judges including the author, examining over 900 different sentences (31 sentences pivot-translated by 32 different languages) in order to come up with a reasonable set of evaluations to rank the languages accordingly. For the study and the results, please see Section 5.2.7.

7.2.5 Comparison of nine sentence level paraphrase algorithms on precision and recall

In Section 5.3, we compare nine different sentence-level transforms, six from our own set and three statistical translation baselines using slightly different chains of translations. Fifty random sentences from the Tatoeba project were used producing 381 different paraphrases by all the transforms.

The 381 sentences were also scored by three judges using the same criteria as the translation study in 5.2.7. The results for each transform are compared on precision and recall They are outlined in Table 28 and graphed in Figure 22.

7.2.6 Summary and self-assessment of additional contributions

We present these additional contributions that are in most cases novel and have research validity of their own. We built these systems and conducted the overwhelming majority of these experiments out of necessity as stepping stones to assessing the main

five hypotheses. Since we did not have explicit hypotheses stated for these additional contributions, we do not conduct the self-assessment along the same rigid scale that we used in Section 7.1.6. This self assessment is a more general statement of significance.

Table 59. Summary and self-assessment of additional contributions

Contribution	Findings	Self-assessment
(7.2.1) Extensible, modular, heterogeneous style processing system	System prototype built according to plan in Figure 2 and Figure 3, complete with all use cases stated in Section 1.3	successfully built prototype combining stylistic classification and generation theories
(7.2.2) Style marker taxonomy	Published 5-layer taxonomy representing most style markers encountered in literature (See Sections 3.4, 3.5)	we assess this to be one of the largest and most comprehensive and most organized collection of style markers in literature
(7.2.3) Evaluation of over 500 markers	502 markers were evaluated on AAAC corpus and ranked using three performance metrics (Section 3.6)	among the largest comparison studies of individual style markers in literature using one of the most diverse and eclectic English corpora (AAAC)
(7.2.4) Evaluating Microsoft Translation languages for paraphrasing	Over 900 paraphrased sentences produced by monolingual pivot translation, evaluated by two judges resulting in a performance evaluation of 32 Microsoft Translation system languages (Section 5.2.7)	novel ranking of languages for paraphrasing using statistical online translation system
(7.2.5) comparison of nine sentence-level paraphrase algorithms on precision and recall	Compared our 6 sentence level transforms with 3 translation baselines (Section 5.3)	diverse comparison of sentence level transforms

7.3 MAIN HYPOTHESIS EVALUATION

Our main hypothesis was the following:

Style processing, the automatic machine controlled analysis, recognition, classification, shifting and transformation of natural language styles, is possible.

Taken as individual as distinct tasks, the literature has many examples of *machine analysis, recognition* and *classification* of styles. These problems have been worked on for a long time under various conceptions of style. *Style shifting* and *transformation* have much less following at the moment, but related activity in style-rich NLG does exist. What makes our contribution unique, we feel, is not the novelty of the individual components but the interaction between them required to perform operations in an intelligent style processing system that we describe in Section 1.3. Figure 35 demonstrates the interaction of the concepts inherent in our presentation of the style processing system.

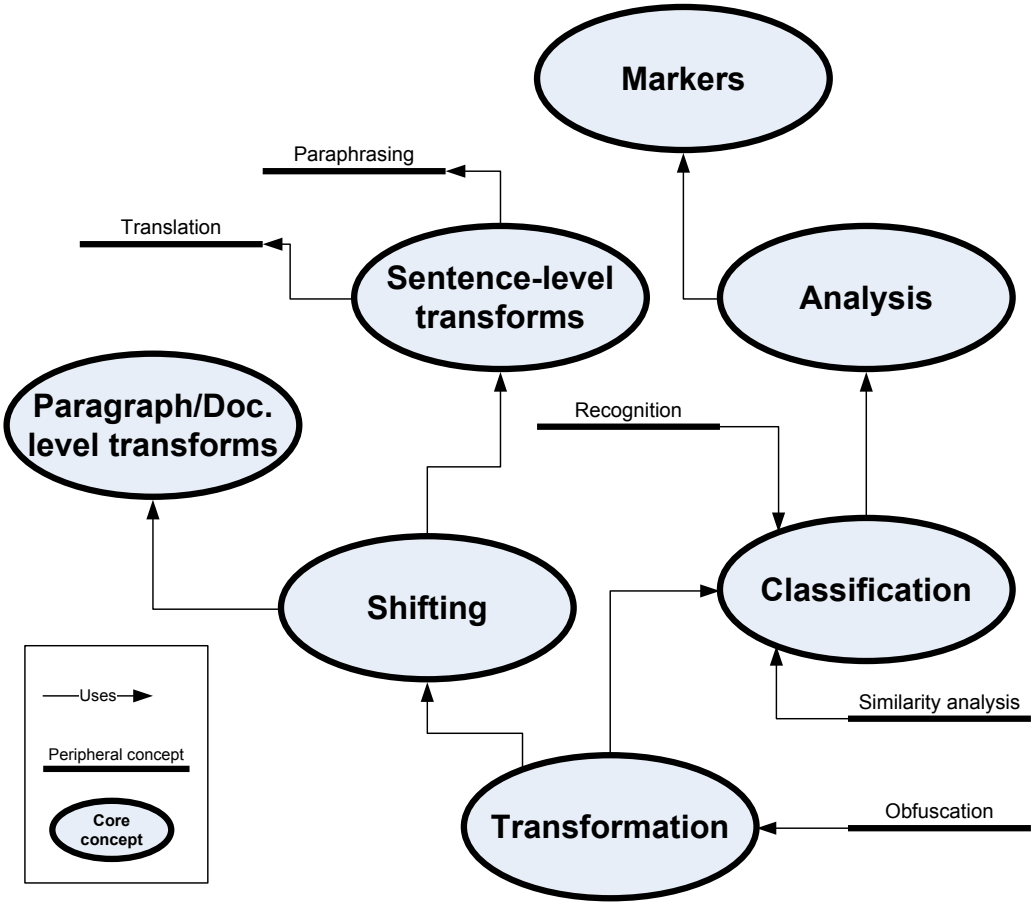

Figure 31. Concepts in style processing system

Five specific hypotheses were listed to strengthen the scientific merits of our work as well as to validate our model of style and style transformation against human understanding of these concepts (Table 56). Although, there is variation in the level of supporting evidence for each hypothesis, we have provided proof for each of them above the baseline. In addition, we have made other contributions (Table 58) that are related to our main work.

During the course of this thesis, scientific experiments were devised and conducted which both provided support for our assumptions, and also exposed important shortcomings. Ultimately, the proof for the main hypothesis, as stated, comes from the fact that we actually built a prototype of a system that does accomplish style processing operations, and further, we evaluated these operations using scientific experiments and user studies and found significant level of support for them.

APPENDIX A: MARKER LIBRARY

In this appendix we present our marker library. **530** markers are used in most experiments of this document, but **754** potential marker instances are currently extractable under these parameters, even though some (like n-grams) can easily be extended to larger numbers. Table 59 lists the appropriate codes for the category-family combinations, the first two layers of the taxonomy hierarchy discussed in Section 3.4. Table 60 lists the markers themselves along with counts of different parameter sets and summary statistics. In most cases, the "5" count in summary statistics refers to: maximum, minimum, mean, median, and variance.

Table 60. Category-family codes for marker library

code	category	Family
A01	lengths	Character
A02	lengths	Word
A03	lengths	Sentence
A04	lengths	Paragraph
A05	lengths	Syllable
A06	lengths	Number
A07	lengths	Vowel
A08	lengths	Punctuation
A09	lengths	Symbol
B01	n-grams	Character
B02	n-grams	Word
B03	n-grams	POS
C01	words	most-frequent
C02	words	least frequent
C03	words	POS ratios
C04	words	Lists
D01	readability	readability measures

D02	readability	Complexity
D03	readability	Voice
D04	readability	diction rules

Table 61. Marker library of 754 potential and 530 used marker instances

category-family code	Marker	param-eters	summary statistics	potential instances	total used
A01	characters per document	2	5	10	0
A01	characters per paragraph	2	5	10	5
A01	characters per sentence	2	5	10	5
A01	characters per word	2	5	10	5
A02	capitalized words per document	1	5	5	5
A02	capitalized words per paragraph	1	5	5	5
A02	capitalized words per sentence	1	5	5	5
A02	capitalized words per word	1	5	5	0
A02	type to token ratio - document	2	5	10	0
A02	type to token ratio – paragraph	2	5	10	0
A02	words per document	2	5	10	5
A02	words per paragraph	2	5	10	5
A02	words per sentence	2	5	10	5
A03	sentences per document	1	5	5	5
A03	sentences per paragraph	1	5	5	5
A04	footer paragraphs per document	1	5	5	5
A04	paragraphs per document	1	5	5	5
A04	title paragraphs per document	1	5	5	5
A05	syllables per paragraph	1	5	5	5
A05	syllables per sentence	1	5	5	5
A05	syllables per word	1	5	5	5
A06	numerics per document	1	5	5	5
A06	numerics per paragraph	1	5	5	5

A06	numerics per sentence	1	5	5	5
A06	numerics per word	1	5	5	0
A07	Vowels per sentence	1	5	5	5
A07	Vowels per word	1	5	5	5
A08	punctuations per document	1	5	5	5
A08	punctuations per paragraph	1	5	5	5
A08	punctuations per sentence	1	5	5	5
A08	punctuations per word	1	5	5	0
B01	character 1-grams	2	5	10	5
B01	character 2-grams	2	5	10	5
B01	character 3-grams	2	5	10	5
B01	character 4-grams	2	5	10	5
B01	character 5-grams	2	5	10	5
B02	Word 1-grams	2	5	10	5
B02	Word 2-grams	2	5	10	5
B02	Word 3-grams	2	5	10	5
B02	Word 4-grams	2	5	10	5
B02	Word 5-grams	2	5	10	5
B03	POS 1-grams	1	5	5	5
B03	POS 2-grams	1	5	5	5
B03	POS 3-grams	1	5	5	5
B03	POS 4-grams	1	5	5	5
B03	POS 5-grams	1	5	5	5
C01	Frequencies of the top 1000 English words	2	5	10	5
C01	Frequencies of the top 500 words	2	5	10	5
C01	top 1000 weighted inversely by popularity	2	5	10	5
C02	counts of hapax legomena	8	1	8	8
C03	adjectives to words	1	1	1	1
C03	adverbs to words	1	1	1	1
C03	all others to words	1	1	1	1
C03	determiners to words	1	1	1	1
C03	nouns to words	1	1	1	1
C03	past tense verbs to verbs	1	1	1	1
C03	plural nouns to nouns	1	1	1	1
C03	pronouns to words	1	1	1	1

C03	proper nouns to nouns	1	1	1	1
C03	VBZ verbs to verbs	1	1	1	1
C03	verbs to words	1	1	1	1
C04	freq. of 10 function words from [Tay87]	10	1	10	0
C04	frequencies of top 100 non-commons	100	1	100	100
C04	fuzzy quantifications per document	3	5	15	15
C04	fuzzy quantifications per paragraph	3	5	15	15
C04	fuzzy quantifications per sentence	3	5	15	15
C04	possible English misspellings / document	1	5	5	5
C04	possible English misspellings / paragraph	1	5	5	5
C04	possible English misspellings / sentence	1	5	5	5
C04	possible English misspellings / word	1	5	5	5
C04	rejected words per document	1	5	5	5
C04	rejected words per paragraph	1	5	5	5
C04	rejected words per sentence	1	5	5	5
C04	rejected words per word	1	5	5	0
C04	stop words per document	1	5	5	5
C04	stop words per paragraph	1	5	5	5
C04	stop words per sentence	1	5	5	5
C04	stop words per word	1	5	5	5
C04	Unix dictionary hits per document	2	5	10	5
C04	Unix dictionary hits per paragraph	2	5	10	5
C04	Unix dictionary hits per sentence	2	5	10	5
C04	Unix dictionary hits per word	2	5	10	5
D01	Automated Readability Index	1	1	1	1
D01	Coleman Liau Index	1	1	1	1
D01	Flesch Reading Ease	1	1	1	1

D01	Flesch-Kinkaid Grade	1	1	1	1
D01	Gunning Fog Index	1	1	1	1
D01	Lix formula	1	1	1	0
D01	SMOG index	1	1	1	1
D02	complex words per document	1	5	5	5
D02	complex words per paragraph	1	5	5	5
D02	complex words per sentence	1	5	5	5
D02	complex words per word	1	5	5	0
D02	Honore's R	1	1	1	0
D02	parser phrase count per sentence	1	5	5	0
D02	quoted segments per paragraph	1	5	5	0
D02	quoted segments per sentence	1	5	5	0
D02	Simpson's rule	1	1	1	0
D02	synsets per document	1	5	5	5
D02	synsets per paragraph	1	5	5	5
D02	synsets per sentence	1	5	5	5
D02	synsets per word	1	5	5	5
D02	syntactic depth per sentence	1	5	5	0
D02	Yule's K	1	1	1	0
D03	active voice sentences / total	2	5	10	0
D03	passive voice sentences / total	2	5	10	0
D04	breadth of rules per document	1	5	5	5
D04	cliché's per document	1	5	5	5
D04	diction variations per document	1	5	5	5
D04	new sent. Suggestions per document	1	5	5	5
D04	omit suggestions per document	1	5	5	5
D04	replacement suggestions per document	1	5	5	5
D04	total diction events per document	1	5	5	5

APPENDIX B: MARKER GLOSSARY FROM [KHO11]

We present the glossary of the markers used in [Kho11]. The set differs slightly from our current library (Appendix A), but was referenced several times in chapters 3 and 4. This is a partial listing of the full 502-marker instance set that was used in [Kho11].

Table 62. Partial listing of markers used in [Kho11] experiments

Code	Marker instance
0	Automated Readability Index (ARI)
2	Flesch-Kinkaid Grade index (FKG)
7	ratio of adjectives to all words
8	ratio of adverbs to all words
9	ratio of other words (other than noun, verb, adjective and adverb) to all
11	ratio of plural nouns to all nouns
12	ratio of past tense verbs to all verbs
13	SMOG index
14	frequency of most common character
19	max. character bigram frequency
31	avg. frequency of top 300 character 4-grams
33	variance of top 300 character 4-grams
39	max. characters per paragraph
40	min. characters per paragraph
44	max. characters per sentence
47	med. characters per sentence
57	med. complex words per document
59	max. complex words per paragraph
61	avg. complex words per paragraph
72	med. number of unique GNU Diction rules applicable per document
79	max. GNU Diction new sentence suggestions
92	med. total number of diction suggestions per document
152	med. number of possible misspellings per document
169	max. numerics per document

172	med. numerics per document
176	avg. numerics per paragraph
210	min. paragraphs per document
211	avg. paragraphs per document
212	med. paragraphs per document
217	med. punctuations per document
226	avg. punctuations per sentence
228	variance of punctuations per sentence
229	max. rejected words per document
230	min. rejected words per document
232	med. rejected words per document
234	max. rejected words per paragraph
236	avg. rejected words per paragraph
244	max. sentences per document
245	min. sentences per document
257	med. stop words per document
262	med. stop words per paragraph
288	variance of syllables per word
290	min. synsets per document
292	med. synsets per document
298	variance of synsets per paragraph
301	avg. synsets per sentence
310	min. Unix dictionary hits
333	variance of vowels per sentence
369	max. words per paragraph
396	frequency of the word "go"
410	frequency of the word "come"
411	frequency of the word "made"
412	frequency of the word "may"
417	frequency of the word "little"
436	frequency of the word "right"
456	frequency of the word "must"
459	frequency of the word "turn"
479	frequency of the word "animal"
480	frequency of the word "house"
484	max. freq. of the most frequent English word
494	ratio of total hapax legomena to all words

APPENDIX C: INDIVIDUAL ALGORITHM-EVENT-SET PERFORMANCE IN JGAAP

These results were obtained using JGAAP [Juo09] version 3.0. The table below represents 100 experiments on the AAAC [Juo04] problem A corpus. All supported event sets and parameters for distance metrics "RN cross entropy" and "Naïve Bayes" are represented. The canonization settings are "strip punctuation" and "normalize white space" for all experiments. The green shadings in a cell indicate a document in that column was correctly attributed to its author.

Table 63. AAAC problem A experiments using JGAAP

Prob. A	Documents:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Event Set	**Parameters**													
[RN cross entropy]														
char.	std	3	3	6	5	6	3	6	3	3	6	3	6	13
	avg	1	13	11	8	10	12	9	9	11	10	10	10	9
	max	7	6	11	8	10	12	9	1	11	10	11	10	9
	min	3	3	6	5	6	3	6	3	3	6	3	6	13
	reverse	10	8	11	10	10	9	9	1	11	10	11	10	9
char.	std	3	3	3	3	7	3	3	3	3	13	3	3	13
bigrams	avg	1	7	7	7	11	1	1	1	2	10	1	10	11
	max	10	7	11	9	11	1	8	4	11	4	11	10	4
	min	3	4	3	3	7	3	3	3	3	13	3	3	13
	reverse	1	4	11	1	1	1	1	1	1	4	1	1	1
char.	std	3	13	3	7	3	3	13	3	3	13	3	13	7
trigrams	avg	1	13	1	1	1	1	1	1	1	4	1	1	1
	max	4	13	4	11	4	1	1	4	5	4	4	4	4
	min	3	1	3	7	3	3	13	3	3	13	3	13	7
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
char.	std	7	13	3	7	13	3	13	7	3	13	13	13	7

4-grams	avg	1	13	1	1	1	1	1	1	5	1	1	1	1
	max	4	13	4	10	4	8	1	4	5	4	5	4	4
	min	7	1	3	7	13	3	13	7	3	13	13	13	7
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
Words	std	12	2	10	4	10	12	8	1	5	4	6	2	9
	avg	4	4	4	4	4	4	4	4	5	4	1	4	4
	max	6	1	4	4	4	2	8	4	5	4	6	9	9
	min	1	2	1	4	10	1	1	1	5	4	1	1	4
	reverse	1	1	1	4	1	1	1	1	5	1	1	1	4
Word	std	9	13	10	10	10	12	10	1	5	4	10	2	4
bigrams	avg	5	5	5	5	5	5	5	5	5	5	5	5	5
	max	9	1	11	10	9	4	10	1	5	5	1	4	9
	min	1	13	1	1	1	1	1	1	5	1	1	1	1
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
Word	std	9	9	9	7	10	12	5	1	5	4	6	12	9
trigrams	avg	5	5	5	5	5	5	5	5	5	5	5	5	5
	max	8	1	9	7	5	5	5	2	5	5	5	11	5
	min	1	9	1	1	1	1	1	1	1	1	1	1	1
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
Word	std	9	13	4	7	5	12	5	2	9	10	6	11	9
4-grams	avg	5	5	5	5	5	5	5	5	5	5	5	5	5
	max	1	1	4	7	5	1	5	2	5	5	5	11	1
	min	1	13	1	1	1	1	1	1	1	1	1	1	1
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
Word	std	2	13	6	7	6	6	12	6	12	7	12	6	7
lengths	avg	2	13	6	7	9	13	8	1	13	1	2	1	7
	max	2	13	6	7	6	13	9	1	13	1	2	11	7
	min	1	1	1	1	1	1	1	1	1	1	1	1	1
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
Syll.	std	6	3	2	1	6	12	2	1	5	1	3	2	7
/ word	avg	6	1	6	1	6	2	1	1	11	1	1	6	1
	max	6	8	6	1	11	12	9	7	11	1	10	6	1
	min	1	1	1	1	1	1	1	1	1	1	1	1	1
	reverse	1	1	1	1	1	1	1	1	1	1	1	1	1
[Naïve Bayes classifier events]														

char.	std	3	3	13	3	10	3	5	6	5	3	3	6	13
	avg	3	3	13	3	10	3	5	6	5	3	3	6	13
	max	3	3	13	3	10	3	5	6	5	3	3	6	13
	min	3	3	13	3	10	3	5	6	5	3	3	6	13
	reverse	3	3	13	3	10	3	5	6	5	3	3	6	13
char.	std	8	8	8	3	7	12	7	6	8	13	4	4	9
bigrams	avg	8	8	8	3	7	12	7	6	8	13	4	4	9
	max	8	8	8	3	7	12	7	6	8	13	4	4	9
	min	8	8	8	3	7	12	7	6	8	13	4	4	9
	reverse	8	8	8	3	7	12	7	6	8	13	4	4	9
char.	std	9	13	13	11	9	3	7	11	10	7	10	3	6
trigrams	avg	9	13	13	11	9	3	7	11	10	7	10	3	6
	max	9	13	13	11	9	3	7	11	10	7	10	3	6
	min	9	13	13	11	9	3	7	11	10	7	10	3	6
	reverse	9	13	13	11	9	3	7	11	10	7	10	3	6
char.	std	11	13	2	8	6	13	12	1	12	3	8	4	12
4-grams	avg	11	13	2	8	6	13	12	1	12	3	8	4	12
	max	11	13	2	8	6	13	12	1	12	3	8	4	12
	min	11	13	2	8	6	13	12	1	12	3	8	4	12
	reverse	11	13	2	8	6	13	12	1	12	3	8	4	12
Words	std	6	7	6	11	7	10	13	6	7	6	6	8	2
	avg	6	7	6	11	7	10	13	6	7	6	6	8	2
	max	6	7	6	11	7	10	13	6	7	6	6	8	2
	min	6	7	6	11	7	10	13	6	7	6	6	8	2
	reverse	6	7	6	11	7	10	13	6	7	6	6	8	2
Word	std	4	12	12	7	12	12	1	4	1	13	6	2	7
bigrams	avg	4	12	12	7	12	12	1	4	1	13	6	2	7
	max	4	12	12	7	12	12	1	4	1	13	6	2	7
	min	4	12	12	7	12	12	1	4	1	13	6	2	7
	reverse	4	12	12	7	12	12	1	4	1	13	6	2	7
Word	std	3	10	9	7	10	12	7	10	9	5	5	11	10
trigrams	avg	3	10	9	7	10	12	7	10	9	5	5	11	10
	max	3	10	9	7	10	12	7	10	9	5	5	11	10
	min	3	10	9	7	10	12	7	10	9	5	5	11	10
	reverse	3	10	9	7	10	12	7	10	9	5	5	11	10
Word	std	1	1	1	1	5	12	1	10	1	1	1	11	1

4-grams	avg	1	1	1	1	5	12	1	10	1	1	1	11	1
	max	1	1	1	1	5	12	1	10	1	1	1	11	1
	min	1	1	1	1	5	12	1	10	1	1	1	11	1
	reverse	1	1	1	1	5	12	1	10	1	1	1	11	1
Word	std	2	3	2	3	3	2	3	12	3	12	2	3	3
lengths	avg	2	3	2	3	3	2	3	12	3	12	2	3	3
	max	2	3	2	3	3	2	3	12	3	12	2	3	3
	min	2	3	2	3	3	2	3	12	3	12	2	3	3
	reverse	2	3	2	3	3	2	3	12	3	12	2	3	3
Syll.	std	3	5	3	5	3	3	5	5	3	8	5	3	5
/ word	avg	3	5	3	5	3	3	5	5	3	8	5	3	5
	max	3	5	3	5	3	3	5	5	3	8	5	3	5
	min	3	5	3	5	3	3	5	5	3	8	5	3	5
	reverse	3	5	3	5	3	3	5	5	3	8	5	3	5
	Correct Authors:	3	13	11	7	10	12	8	1	5	4	6	2	9

APPENDIX D: LOG FILE FROM A TYPICAL TRANSFORMATION OPERATION

This file is provided for informational purposed and to demonstrate the incremental nature of style transformation. It is a very high level log reporting only the most encapsulating events of the system. Line numbers are provided on the left side.

The system sets up the problem in **lines 1-3**. The source file (file to be modified) is "B4was2-0" which in our shorthand means "AAAC problem B version 4, and author is Author 02". The target is Author01, so the transform attempts to get the source close to the target so it can be reclassified as such.

In **line 48** the system reports that this reclassification has been accomplished, and the text now is classified as "Author01". However, as we find, in most situations, being "just barely" classified as the intended target is not good enough. At least from the perspective of human readers it appears borderline at best. Thus we instruct the system to continue making changes, and minimizing the distance to the target, until no closer step is possible. This finally occurs in **line 110**. There the system reports "one cycle" (every sentence examined) has passed without a single replacement being adopted, so the system halts.

```
1.  Starting Transform Log for file B4was2-0 Label: X Target: Author01

2.  initial attribution: Author02
```

3. **initial distance to target: 0.336097259666**
4. Transforms class variable setup
5. Paragraph] no index for target Paragraph stats, using fake stats to initiate
6. Paragraph] no change
7. new distance to target: 0.325361567508
8. new sentence adopted: The irony of the situation is the sympathy our nation has for Monica Lewinski. => The irony of the state is the sympathy our nation has for Monica Lewinski.
9. new distance to target: 0.326574947892
10. new sentence rejected: Her story has portrayed her as somewhat of a victim, while President Clinton is being portrayed as the abuser. => Her story has portrayed her as pretty of a victim, while President Clinton is being portrayed as the abuser.
11. new distance to target: 0.322532053568
12. new sentence adopted: Yet she knew very well that he was a married man and that what they were doing could have ruined his career and marriage. => Yet she knew fine that he was a married man and that what they were doing could have ruined his career and marriage.
13. new distance to target: 0.317519760416
14. new sentence adopted: She has been a topic on several daily talk shows, and she has made many appearances on late night television. => She has been a topic on several daily talk shows, and many appearances have been made by her on late night television.
15. new distance to target: 0.338218444363
16. new sentence rejected: The exclusive personal interviews discussing her dramatic situation are countless. => The exclusive personal interviews discussing her dramatic situation is innumerable.
17. new distance to target: 0.312813130346
18. new sentence adopted: She has been the object of Saturday Night Live skits, and the focus on other comical shows. => She has been the object of Saturday Night Live playings, and the focus on other comical shows.
19. new distance to target: 0.311013156544
20. new sentence adopted: She has even designed her own collection of handbags. => She was even designed by her own collection of handbags.
21. new distance to target: 0.311724019954
22. new sentence rejected: To say she has been over-publicized is an extreme understatement. => To tell she has been over-publicized is an extreme understatement.
23. new distance to target: 0.295498306796
24. new sentence adopted: In reality while gaining a negative reputation from the affair she has actually benefited, at least financially. => In reality while gaining a negative honor from the affair she has actually benefited, at least financially.
25. writing new version of file... B4was2-00

26. End of cycle 0 currentLabel: Author02
27. distance to target: 0.295498306796

28. Paragraph] no change
29. new distance to target: 0.27910739526
30. new sentence adopted: The irony of the state is the sympathy our nation has for Monica Lewinski. => The irony of the State is the sympathy, our nation has for Monica Sent.
31. new distance to target: 0.278286205603

32. new sentence adopted: Her story has portrayed her as somewhat of a victim, while President Clinton is being portrayed as the abuser. => Her story has portrayed her as pretty of a victim, while President Clinton is being portrayed as the abuser.
33. new distance to target: 0.277248589622
34. new sentence adopted: Yet she knew fine that he was a married man and that what they were doing could have ruined his career and marriage. => Yet she did it fine that he was a married man and that what they were doing could have ruined his career and marriage.
35. new distance to target: 0.264594265086
36. new sentence adopted: She has been a topic on several daily talk shows, and many appearances have been made by her on late night television. => She has been a topic on several daily talk shows, and many appearances have been made by her on late night TV.
37. new distance to target: 0.278768074617
38. new sentence rejected: The exclusive personal interviews discussing her dramatic situation are countless. => The exclusive personal interviews discussing her dramatic situation is innumerable.
39. new distance to target: 0.263907004707
40. new sentence adopted: She has been the object of Saturday Night Live playings, and the focus on other comical shows. => She has been the object of Saturday Night Live playings, and the center other comical shows.
41. new distance to target: 0.263412789035
42. new sentence adopted: She was even designed by her own collection of handbags. => She was even designed by her have collection of handbags.
43. new distance to target: 0.264109303036
44. new sentence rejected: To say she has been over-publicized is an extreme understatement. => To tell she has been over-publicized is an extreme understatement.
45. new distance to target: 0.251903016916
46. new sentence adopted: In reality while gaining a negative honor from the affair she has actually benefited, at least financially. => while gaining a negative honor from the affair she has benefited, at least financially.
47. writing new version of file... B4was2-01
48. Xform Success! Text has been reclassified as Author01

49. Paragraph] no change
50. new distance to target: 0.250547818476
51. new sentence adopted: The irony of the State is the sympathy, our nation has for Monica Sent. => The irony of the State is sympathy, our nation to Monica sent.
52. new distance to target: 0.257591051058
53. new sentence rejected: Her story has portrayed her as pretty of a victim, while President Clinton is being portrayed as the abuser. => Her story has portrayed her as pretty of a victim, while President Clinton is being depicted as the abuser.
54. new distance to target: 0.250623104634
55. new sentence rejected: Yet she did it fine that he was a married man and that what they were doing could have ruined his career and marriage. => Yet she did it fine that he was a married man and that what they were doing could have ruined his business and marriage.
56. new distance to target: 0.251579267756
57. new sentence rejected: She has been a topic on several daily talk shows,

and many appearances have been made by her on late night TV. => She has been a topic on several daily talk shows, and many appearances have been made by her on later night TV.
58. new distance to target: 0.261881733073
59. new sentence rejected: The exclusive personal interviews discussing her dramatic situation are countless. => The exclusive personal interviews discussing her dramatic situation is innumerable.
60. new distance to target: 0.249537095835
61. new sentence adopted: She has been the object of Saturday Night Live playings, and the center other comical shows. => She has been the object of Saturday Night Live playings, and the heart other comical shows.
62. new distance to target: 0.243898011368
63. new sentence adopted: She was even designed by her have collection of handbags. => She was even designed by her have group of handbags.
64. new distance to target: 0.244639988363
65. new sentence rejected: To say she has been over-publicized is an extreme understatement. => To tell she has been over-publicized is an extreme understatement.
66. new distance to target: 0.242226991158
67. new sentence adopted: while gaining a negative honor from the affair she has benefited, at least financially. => While gaining a negative honor from the thing she has benefited, at least financially.
68. writing new version of file... B4was2-02
69. Xform Success! Text has been reclassified as Author01

70. Paragraph] no change
71. new distance to target: 0.25137288241
72. new sentence rejected: The irony of the State is sympathy, our nation to Monica sent. => The irony of the State is sympathy, our nation to Monica directed.
73. new distance to target: 0.247344652333
74. new sentence rejected: Her story has portrayed her as pretty of a victim, while President Clinton is being portrayed as the abuser. => Her story has portrayed her as pretty of a victim, while President Clinton is being depicted as the abuser.
75. new distance to target: 0.242300709994
76. new sentence rejected: Yet she did it fine that he was a married man and that what they were doing could have ruined his career and marriage. => Yet she did it fine that he was a married man and that what they were doing could have ruined his business and marriage.
77. new distance to target: 0.242957200293
78. new sentence rejected: She has been a topic on several daily talk shows, and many appearances have been made by her on late night TV. => She has been a topic on several daily talk shows, and many appearances have been made by her on later night TV.
79. new distance to target: 0.251030078944
80. new sentence rejected: The exclusive personal interviews discussing her dramatic situation are countless. => The exclusive personal interviews discussing her dramatic situation is innumerable.
81. new distance to target: 0.242957474647
82. new sentence rejected: She has been the object of Saturday Night Live playings, and the heart other comical shows. => She has been the object of Saturday Night Live playings, and the centre other comical shows.
83. new distance to target: 0.247622844431
84. new sentence rejected: She was even designed by her have group of

handbags. => She herself was designed by her, the Group of handbags.
85. new distance to target: 0.242971605484
86. new sentence rejected: To say she has been over-publicized is an extreme understatement. => To tell she has been over-publicized is an extreme understatement.
87. new distance to target: 0.232241554271
88. new sentence adopted: While gaining a negative honor from the thing she has benefited, at least financially. => At the same time, wins a negative honor from the things she has learned, at least financially.
89. writing new version of file... B4was2-03
90. Xform Success! Text has been reclassified as Author01

91. Paragraph] no change
92. new distance to target: 0.238795023583
93. new sentence rejected: The irony of the State is sympathy, our nation to Monica sent. => The irony of the State is sympathy, our nation to Monica directed.
94. new distance to target: 0.235423112019
95. new sentence rejected: Her story has portrayed her as pretty of a victim, while President Clinton is being portrayed as the abuser. => Her story has portrayed her as pretty of a victim, while President Clinton is being depicted as the abuser.
96. new distance to target: 0.232318599162
97. new sentence rejected: Yet she did it fine that he was a married man and that what they were doing could have ruined his career and marriage. => Yet she did it fine that he was a married man and that what they were doing could have ruined his business and marriage.
98. new distance to target: 0.232614361216
99. new sentence rejected: She has been a topic on several daily talk shows, and many appearances have been made by her on late night TV. => She has been a topic on several daily talk shows, and many appearances have been made by her on later night TV.
100. new distance to target: 0.238356736559
101. new sentence rejected: The exclusive personal interviews discussing her dramatic situation are countless. => The exclusive personal interviews discussing her dramatic situation is innumerable.
102. new distance to target: 0.232615045187
103. new sentence rejected: She has been the object of Saturday Night Live playings, and the heart other comical shows. => She has been the object of Saturday Night Live playings, and the centre other comical shows.
104. new distance to target: 0.235468750511
105. new sentence rejected: She was even designed by her have group of handbags. => She herself was designed by her, the Group of handbags.
106. new distance to target: 0.232939343251
107. new sentence rejected: To say she has been over-publicized is an extreme understatement. => To tell she has been over-publicized is an extreme understatement.
108. new distance to target: 0.236591990185
109. new sentence rejected: At the same time, wins a negative honor from the things she has learned, at least economically. => At the same time, wins a negative honor from the situations she has learned, at least economically.
110. EXIT: one cycle without any adopted changes... ending program.

BIBLIOGRAPHY

[Abb05] A. Abbasi and H. Chen, "Identification and comparison of extremist-group Web forum messages using authorship analysis," IEEE Intelligent Systems 20, 2005, pp.67-75.

[Abb06] A. Abbasi and H. Chen, "Visualizing authorship for identification", in Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics, San Diego, Ca. 2006.

[Abb08] A. Abbasi and H. Chen, "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace," ACM Transactions on Information Systems, Vol. 26, No. 2, Article 7, March 2008.

[Arg98] S. Argamon, M. Koppel and G. Avneri, "Routing documents according to style," in Proceedings of the 1st International Workshop of innovative Information. 1998.

[Arg03] S. Argamon, M. Saric, S. Stein, "Style Mining of electronic messages for multiple authorship discrimination: first results.", Proceedings of the 9th ACM SIGKDD, Washington DC., 2003.

[Art11] Article Rewriter, accessed July 2011. http://article-rewriter.appspot.com/

[Baa96] H. Baayen, H. van Haltern, F. Tweedie, "Outside thecave of shadows': using syntactic annotations to enhance authorship attribution," Literary an Linguistic Computing 11, 1996, pp. 121-132.

[Bar01] Regina Barzilay and Kathleen McKeown, "Extracting paraphrases from a parallel corpus," in *Proceedings of ACL*, Toulouse, France, 2001.

[Bib88] Douglas Biber, *Variation across speech and writing*, Cambridge University Press, 1988.

[Bib89] Douglas Biber, "A typology of English texts", *Linguistics* 27 (1989), 3–43.

[Bib98] D. Biber, S. Conrad and R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use,* Chapter 5, Cambridge University Press, 1998.

[Bra97] Richard Bradford, *Stylistics*, part of the The New Critical Idiom series, Routlidge, 1997.

[Bre09] M. Brennan and R. Greenstadt, "Practical attacks against authorship recognition techniques," in *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence* (IAAI), Pasadena, CA, 2009.

[Bri00] Eric Brill, "Part-of-Speech Tagging", in Handbook of Natural Language Processing edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 403-414.

[Brin00] S. Bringsjord, and Ferrucci, Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine, Mahwah, NJ: Lawrence Erlbaum, 2000.

[Bou11] Houda Bouamor, Aurélien Max, Gabriel Illouz, Anne Vilnat, "Web-based validation for contextual targeted paraphrasing," *Proceedings of the ACL Workshop on Monolingual Text-To-Text Generation* (Text-To-Text-2011), Portland, Oregon, USA, 2011.

[Buf21] Comte de Buffon, "Discourse on Style," trans. Rollo Walter Brown, in The Writer's Art, ed. Brown, Harvard University Press, 1921, pp. 285-86. (originally published 1773)

[Car88] Ronald Carter, and Paul Simpson, Language, *Discourse and Literature: An Introductory Reader in Discourse Stylistics*, Routledge, 1988.

[Cim06] Philipp Cimiano, *Ontology Learning and Population from Text*, Springer, 2006.

[Con09] ConceptNet, MIT Common Sense Computing initiative's ConceptNet (v 3) API. http://csc.media.mit.edu/docs/conceptnet/, 2009.

[Cos11] Will Coster and David Kauchak, "Learning to Simplify Sentences Using Wikipedia," *Proceedings of the ACL Workshop on Monolingual Text-To-Text Generation* (Text-To-Text-2011), Portland, Oregon, USA, 2011.

[Cry69] David Crystal and Derek Davy, *Investigating English style*, Indiana University Press, 1969.

[Dag00] Ido Dagan, "Contextual Word Similarity", in *Handbook of Natural Language Processing* edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 459-463.

[DeV01] O. De Vel, A. Anderson, M. Corney and G Mohay, "Mining email content for author identification forensics," ACM SIGMOD Rec. 30, 4, pp. 55-64.

[Dim94] Chrysanne DiMarco, "Stylistic Choice in Machine Translation," AMAT, 1994.

[Eng10] EnglishPage, by Language Dynamics, (www.englishpage.com), accessed Dec. 2010.

[Fak01] N. Fakotakis, E. Stamatatos and G. Kokkinakis. "Computer-based Attribution without Lexical Measures." Computers and the Humanities, Volume 35, Issue 2, May 2001, pp. 193-214

[Fer03] Giacomo Ferrari, "State of the art in Computational Linguistics," in Linguistics Today: Facing a greater Challenge, International Congress of Linguists, John Benjamins Publishing Company, 2003, p 163.

[Fis81] Stanley Fish, "What is stylistics and why are they saying such terrible things about it", in Essays in Modern Stylistics, edited by DC Freeman, Routledge, 1981, pp 53-66.

[Fow26] Henry W. Fowler, A Dictionary of Modern English Usage, Oxford University Press, 1926.

[Fra64] W. N. Francis and H. Kucera, *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*, Department of Linguistics, Brown University, 1964, Providence, Rhode Island, USA.

[Gre94] Gregory Grefenstette, *Explorations in Automatic Thesaurus Discovery*, Kulwar Academic Publishers, 1994.

[Ger00] P. Gervas, "Wasp: Evaluation of different strategies for the automatic generation of Spanish verse," in Proceedings of the AISB00 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science, 2000.

[Gri95] D. Grinberg, J. Lafferty and D. Sleator, "A robust parsing algorithm for link grammars", Carnegie Mellon University Computer Science technical report CMU-CS-95-125, and Proceedings of the Fourth International Workshop on Parsing Technologies, Prague, September, 1995.

[Haa07] Michael Haardt, GNU *diction(1)* PDF manual, accompanying *diction* version 1.11. 2007. http://www.gnu.org/software/diction/diction.html

[Har68] Zellig Harris, *Mathematical Structures of Language*, Wiley, 1968.

[Hei00] George E. Heidorn, "Intelligent Writing Assistance", in *Handbook of Natural Language Processing* edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 181-209.

[Hol95] D. I. Holmes and R. S. Forsyth, "The Federalist Revisited: New Directions in Authorship Attribution. Literary and Linguistic Computing 10(2), 1995, pp. 111–127.

[Hoo99] D. Hoover, *Language and style in The Inheritors*, University Press of America, 1999.

[Hov88] Eduard Hovy, *Generating Natural Language under Pragmatic Constraints*, Lawrence Erlbaum Associates, 1998.

[Iqb08] F. Iqbal, R. Hadjidja, B. Funga, and M. Debbabia, "A novel approach of mining write-prints for authorship attribution in e-mail forensics", *The Proceedings of the Eighth Annual DFRWS Conference, Digital Investigation* Volume 5, Supplement 1, September 2008, S42-S51.

[Ink03] D.Z. Inkpen and G. Hirst, "Near-synonym choice in anatural language generation", in *Proceedings of the International Conference of RANLP*-2003.

[Ink06] Diana Zaiu Inkpen, Olga Feiguina, and Graeme Hirst, "Generating more-positive or more-negative text'', in *Computing Attitude and Affect in Text*, Springer, Dordrecht, The Netherlands, p.187-196. (Selected papers from the *Proceedings of the Workshop on Attitude and Affect in Text, AAAI 2004 Spring Symposium*), Edited by James G. Shanahan, Yan Qu, Janyce Wiebe, 2006.

[Jac97] C. Jacquemin, J. Klavans, and E. Tzoukermann, "Expansion of multi-word terms for indexing and retrieval using morphology and syntax," in proceedings of the 35th Annual Meeting of the ACL, pages 24–31, Madrid, Spain, July 1997.

[Jia97] Jay Jiang and David Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), Taiwan, 1997.

[Juc92] Andreas H. Jucker, *Social Stylistics: syntactic variation in British newspapers*, Walter de Gruiyter, 1992.

[Juo04] P. Juola, "Ad-hoc authorship attribution competition," in Proc. of Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computer and the Humanities (ALLC/ACH 2004), Goteborg, Sweden, June 2004.

[Juo06] P. Juola, *Authorship Attribution*, NOW Publishers, 2006

[Juo09] Patrick Juola, et. al. JGAAP, a Java-based,modular, program for textual analysis, text categorization, and authorship attribution. www.jgaap.com, 2009.

[Juo10] Patrick Juola, "Empirical evaluation of authorship obfuscation using JGAAP," in *Proceedings of the 3rd ACM workshop on Artificial intelligence and security*, AISec, 2010.

[Kac06] G. Kacmarcik, and G. Gamon, "Obfuscating document stylometry to preserve author anonymity," proceedings of the COLING/ACL on Main conference poster sessions 444-451, 2006.

[Kar04] Jussi Karlgren, “The wheres and whyfores for studying text genre computationally,” In Style and Meaning in Language, Art, Music and Design, Washington D.C., 2004. AAAI Symposium series.

[Kes10] Fazel Keshtkar and Diana Inkpen, "A Corpus-based Method for Extracting Paraphrases of Emotion Terms", in *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA, June 2010, pp. 35-44.

[Kes03] Vlado Kesselj et. al .“N gram-based Author Profiles for Authorship Attribution.” in *Proceedings of the Conference Pacific Association for Computational Linguistics*, PACLING'03, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.

[Kho06] Foaad Khosmood and Robert Levinson, “Toward Unification of Source Attribution Processes and Techniques,” IEEE ICMLC, Dalian, China, August 2006.

[Kho08] Foaad Khosmood and Robert Levinson, “Automatic Natural Language Style Classification and Transformation,” BCS Corpus Profiling Workshop, October 2008, London, UK.

[Kho09] Foaad Khosmood and Robert Levinson, "Toward Automated Stylistic Transformation of Natural Language Text," Digital Humanities 09, University of Maryland June 2009.

[Kho10] Foaad Khosmood and Marilyn Walker, "Grapevine: A Gossip Generation System", Foundations of Digital Games 2010, Monterey, California, June 2010.

[Kho10-2] Foaad Khosmood and Robert Levinson, "Automatic Synonym and Phrase Replacement Shows Promise for Style Transformation", IEEE International Conference of Machine Learning and Applications, 2010, Washington, DC, December 2010.

[Kho11] Foaad Khosmood and Robert Levinson, "Taxonomy and Evaluation of Markers for Computational Stylistics", International Conference on Artificial Intelligence, 2011 (ICAI '11), Las Vegas, NV, July 2011.

[Lan98] T. K. Landauer and P. W. Foltz and D. Laham, “Introduction to Latent Semantic Analysis,” 1998, http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

[Lan94] I. Langkilde and K. Knight, “Generation that exploits corpus based statistical knowledge”, in *Proceedings of the ACL-COLING*, 1998.

[Lan02] I. Langkilde-Geary, *A Foundation for the General Purpose Natural Language Generation: Sentence Realization Using Probability Models of Language*. Ph.D. Dissertation, University Southern California, 2002.

[Lav97] Benoit Lavoie and Owen Rambow, "A fast and portable realizer for text generation systems", in *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pp. 265-268, Washing, D.C., 1997.

[Loe96] Dan. Loehr, "An Integration of a Pun Generator with a Natural Language Robot," In: Proceedings of the International Workshop on Computational Humor, Enschede, Netherlands. University of Twente, 1996.

[Lat08] Latent Semantic Analysis resources at University of Colorado, accessed January, 2008. http://lsa.colorado.edu

[Lev66] V. I. Levenshtein, " Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, 1966.

[Li06] J. Li, R. Zheng and H. Chen, "From Fingerprint to Writeprint," Communications of the ACM, 49(4), 76–82. 2006.

[Lik32] Rensis Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology* 140: 1–55. 1932.

[Lut05] Gary Luz, Diane Stevenson, Writer's Digest Grammar Desk Reference, Writer's Digest Books, 2005.

[Luy04] Kim Luyckx and Walter Daelemans, "Shallow text analysis and machine learning for authorship attribution.", In: Computational Linguistics in the Netherlands 2004: selected papers from the Fifteenth CLIN Meeting / van der Wouden Ton [edit.], e.a., Utrecht, LOT, 2005, p. 149-160.

[Mai08] Francois Mairesse and M. A. Walker, "PERSONAGE: Personality Generation for Dialogue", 45th Annual Meeting of the Association for Computational Linguistics, Prague, June 2007.

[Mai10] F. Mairesse and M. A. Walker, "Towards Personality-Based User Adaptation: Psychologically Informed Stylistic Language Generation", *User Modeling and User-Adapted Interaction*, vol. 20, issue 3, 2010.

[McE00] Tony McEnery and Michael Oakes, "Authorship Identification and Computational Stylometry", in *Handbook of Natural Language Processing* edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 548.

[Mel08] Herman Melville, *Moby Dick; or the Whale*, Project Gutenberg E-book #2701, produced December, 2008.

[Men87] T.C. Mendenhall, "The characteristic curves of composition," Science, 11, 1887, pp. 237-249.

[Mer93] TVN Merriam, "Marlowe's hand in *Edward III*", Literary and Linguistic Computing 8, 1993, pp. 59-72.

[Mer96] TVN Merriam, "Marlowe's hand in Edward III revisited", Literary and Linguistic Computing 11, 1996, pp. 19-22.

[Mic05] Jason Michelizzi, "Semantic Relatedness Applied to All Words Sense Disambiguation," thesis submitted to the Faculty of Graduate School, University of Minnesota, July 2005.

[Mil90] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, "Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue), 3(4):235–245, 1990.

[Moe01] Lilo Moessner, "Genre, Text Type, Style, Register, Terminological Maze?" European Journal of English Studies, 2001, Vol. 5, No. 2, pp. 131–138

[Mos84] F. Mosteller and DL Wallace, Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Reading, MA: Addison-Wesley.

[MSN11] Micorsoft Web N-Gram service, public Beta program, http://web-ngram.research.microsoft.com/info/ accessed 2011.

[Mur22] John Middleton Murry, *The Problem of Style* (London: Oxford University Press, 1922), p. 77.

[NIS02] NIST, "Automatic Evaluation of Machine Translation Quality using N-gram Co Occurrence Statistics." http://www.nist.gov/speech/tests/mt/doc/ngramstudy.pdf

[NLT09] NLTK, The Natural Language Tool Kit, project home page: http://www.nltk.org/, 2009.

[Pal00] David D. Palmer, "Tokenisation and Sentence Segmentation", in *Handbook of Natural Language Processing* edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 11.

[PEC01] The Plain English Campaign, "A to Z of alternative words", 2001. http://www.plainenglish.co.uk/files/alternative.pdf

[Rao00] J. Rao and P. Rohatgi, "Can pseudonomity really guarantee privacy," Proccedings of the 9th USENIX Security Symposium 9:7-7, 2000.

[Rij79] C. J. van Rijsbergen, *Information Retrieval*, London, Butterworths, 1979.

[Ris94] Matti Rissanen, "The Helsinki Corpus of English Texts" in *Corpora Across the Centuries, Proceedings of the First International Colloquium on English Diachronic Corpora*, St. Catharine's College Cambridge, 25–27 March 1993, eds. Merja Kyto, Matti Rissanen and Susan Wright (Amsterdam/Atlanta, GA: Rodopi, 1994), 73–79, pp. 76–7.

[Rud03] J. Rudman, "Cherry Picking in Nontraditional Authorship Attribution Studies," CHANCE, vol. 16, No. 1, 2003.

[Sid10] Advaith Siddharthan, "Complex lexico-syntactic reformulation of sentences using typed dependency representations". In Proceedings of the 6th International Natural Language Generation Conference (INLG 2010), pages 125-133, Dublin, Ireland, 2010.

[Sci05] Scigen - an automatic cs paper generator. 2005, http://pdos.csail.mit.edu

[Sim04] John Simpson, *Stylistics: A Resource Book for Students*, Routledge, 2004.

[Sle93] Daniel Sleator and Davy Temperly. "Parsing English with a Link Grammar". Third International Workshop on Parsing Technologies, 1993.

[Str18] William Strunk, *The elements of style*, Ithaca, N.Y.: Priv. print., 1918.

[Sta09] Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods," in *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[Tat10] Project Tatoeba English language sentence corpus (about 150,000 sentences). accessed from tatoeba.org, September 2010.

[Tay87] G. Taylor, “The canon and chronology of Shakespeare’s plays,” in S Wells, G Tayor, editors. *William Shakespeare: A Textual Companion*. Oxford:, Clarendon Press, pp 69-1145.

[Tou03] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”, in *Proceedings of HLT-NAACL* 2003, pp. 252-259.

[Tse04] Georgios Tserdanelis and Wai Yi Peggy Wong, (editors), *Language Files: Materials for an Introduction to Language and Linguistics*, 9th edition, The Ohio State University Press, Columbus, p. 365, 2004.

[Van03] Piet Van Sterkenburg, Editor, *Linguistics Today: Facing a greater Challenge*, International Congress of Linguists, John Benjamins Publishing Company, 2003.

[Wal02] M. A. Walker, and O. Rambow, "Spoken Language Generation", Computer Speech and Language Special Issue on Spoken Language Generation, July, 2002.

[Wal80] Jane Walpole, “Style as Option,” College Composition and Communication, vol. 31, No. 2, Recent Work in Rhetoric: Discourse Theory, Invention, Arrangement, Style, Audience, (May, 1980), pp. 205-212.

[Whi04] Casey Whitelaw and Shlomo Argamon, “Systemic Functional Features in Stylistic Text Classification”, AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design, October 2004.

[Wor08] WordNet at Princeton University Cognitive Science Library, http://wordnet.princeton.edu, accessed 9/2008.

[Wub11] Sander Wubben, Erwin Marsi, Antal van den Bosch and Emiel Krahmer, “Comparing Phrase-based and Syntax-based Paraphrase Generation,” *Proceedings of the ACL Workshop on Monolingual Text-To-Text Generation* (Text-To-Text-2011), Portland, Oregon, USA, 2011.

[Yul38] G.U. Yule, “On sentence length as a statistical characteristic of style in prose,” Biometrika 30, 1938, pp. 363-390.

[Yul44] G.U. Yule, *Statistical study of literary vocabulary*, Cambridge University Press, 1944, Cambridge, UK.

[Zha07] Shiqi Zhao, Ting Liu, Xincheng Yuan, Sheng Li, and Yu Zhang, "Automatic acquisition of context specific lexical paraphrases," in *Proceedings of IJCAI*, Hyderabad, India, 2007.

[Zha09] Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Lib "Application-driven statistical paraphrase generation, in *Proceedings of the Joint ACL-IJCNLP*, Singapore, 2009.

[Zha10] Shiqi Zhao, Haifeng Wang, Ting Liu, , and Sheng Li, "Leveraging multiple MT engines for paraphrase generation," in *Proceedings of COLING*, Beijing, China, 2010.

[Zhe06] R. Zheng, J. Li, Z. Huang and H. Cohen, "A framework for authorship analysis of online messages: Writing style features and techniques," Journal of American Society for Information, Science and Technology, 57, 3, 2006, pp. 378-393.

[Zhu10] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych, "A monolingual tree-based translation model for sentence simplification," In Proceedings of COLING, Beijing, China, 2010.