Automatic Natural Language Style Classification and Transformation

Foaad Khosmood and Robert A. Levinson
University of California Santa Cruz
Department of Computer Science, 1156 High St, Santa Cruz, CA 95064, USA
foaad@soe.ucsc.edu, levinson@soe.ucsc.edu

Style is an integral part of natural language in written, spoken or machine generated forms. Humans have been dealing with style in language since the beginnings of language itself, but computers and machine processes have only recently begun to process natural language styles. Automatic processing of styles poses two interrelated challenges: classification and transformation. There have been recent advances in corpus classification, automatic clustering and authorship attribution along many dimensions but little work directly related to writing styles directly and even less in transformation. In this paper we examine relevant literature to define and operationalize a notion of "style" which we employ to designate style markers usable in classification machines. A measurable reading of these markers also helps guide style transformation algorithms. We demonstrate the concept by showing a detectable stylistic shift in a sample piece of text relative to a target corpus. We present ongoing work in building a comprehensive style recognition and transformation system and discuss our results.

style, natural language processing, artificial intelligence, corpus classification, style recognition, style transformation

1. INTRODUCTION

For millennia, human scholars have been studying, debating "style"; authors have been writing and changing written styles and critics have been commenting on those styles. The concept of style, as well as detection of a particular instance of a style, and even rewriting of the same text in a different style seems like a trivial task to most human experts. We are now interested in automating some of this process to perform classification and transformation with minimal human interference and knowledge.

Style classification and transformation have numerous applications in information retrieval (IR), natural language processing (NLP), human computer interaction (HCI), and interactive entertainment. Robust style classification can lead to an entirely new dimension in searching. Not only are new search parameters such as style markers and related tolerance levels possible, but search systems could adopt style searching based on example of an input text. Style comparison techniques could provide for more robust and descriptive plagiarism detection and digital forensics. Individuals, too, can use style transformation software to obfuscate and alter their own style of writing for online privacy reasons. In HCI, richer and more customizable user help interfaces are possible. Texts for such interfaces could gradually adapt to a particular user's style and facilitate easier understanding. Natural language generation (NLG) systems such as (Gervas 2000), (Goncalo 2007), (Loehr 1996), (Mairesse 2008) and (Scigen 2005) already incorporate some level of style in their generation work, but could enhance their stylistic variability with text-to-text style transformations. In machine translation (MT) (DiMarco 1994), automatic style processing could lead to better and more precise translations. Traditional writers and satirists could use style processing to help compose language in deliberate and memorable styles. In interactive entertainment (Bringsjord 2000), authoring tools for generating game narratives and non-player character (NPC) dialogue could be made much richer, more diverse and more nuanced without significantly increasing a human composer's burden.

It is necessary to examine the very notion of writing "style" as used by humans in different contexts, before we can operationalize it for computational processes.

The manner in which styles can be recognized and used in a transformation system is through detection and manipulation of style markers. These are linguistic and non-linguistic atomic features of written text. We associate one or more of these markers to a particular style. It is also important to examine the literature for other research on these markers.

In the following sections, we first situate style within some neighboring concepts in linguistics and literary studies. In the next section, we present a number of definitions from various fields for “style” and “stylistics.” We examine some academic discussions within the literary studies community on the merit of traditional style and stylistics as disciplines. Much of this discussion centers on differing interpretations of style.

We examine Walpole’s “Style as Option” notion (1980) and its roots in the literature. A later section delves into style markers, or “atoms of style” that can be detected as features by software in the text. Finally we present our design for a comprehensive stylistics analysis system, results from our ongoing experiments, conclusion and future work.

2. GENRE, REGISTER AND TEXT-TYPE

Style as a term has too broad a definition, even when applied only to the written text. We begin our investigation of style by examining some neighboring concepts that sometimes overlap with what is commonly called “style.” Three of these concepts are text-type, genre, and register. The latter term is favored in linguistics at the expense of the less defined “style.”

Moessner (2001) provides 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 (Rissanen 1991) 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” (Moessner 2001) in addition to a catch-all “none of the above” category. Thus (Rissanen 1981) called their classification variable “text type.”

Both text type and genre are non-linguistic classifications, but each of their classes exhibit high correlation to one or more linguistic features. As Moessner 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. (2001)

This gives a definition of “register” as a purely linguistically defined category. So is the term “style” but only as a modifier of genre and text type.

Biber is able to derive purely linguistic text types by aggregating and clustering linguistic features. Biber (1989) identifies 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, Biber produced 8 “text type” categories derived using automatic clustering algorithms.

3. STYLE AND STYLISTICS

Stylistics has been called a conceptual successor to the ancient Greek concept of rhetoric (Bradford 1997) Andreas Jucker (1992) likened “style” to what de Sassure calls “parole” and what Chomsky calls “performance.” Jucker (1992) calls style as 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”. 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. (Jucker 1992)

Other definitions of stylistics exist in the literature.

Walpole calls stylistics an “offshoot of linguistics” which “takes the position” that “style is a deviation from normal language usage” (1980). Style can be contextualized with this definition:

> A style is a consistent and distinguishable tendency to make some [of these] linguistic choices. Style is on the surface level, very obviously detectable as the choices between items in a vocabulary, between types of syntactical constructions, between the various ways a text can be woven from the material it is made of. (Karlgren 2004)

Freeman in his *Essays in Modern Stylistics* called stylistics simply the “application of linguistics to the study of literature (Fish 1981).

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” (2004). George Heidorn of Microsoft is probably closer to the lay definition of “style”, when he defines “style checking” as follows:

> Style checking refers to checking text for errors that a book on good writing style, would discuss, such as overuse of the passive voice. (Heidorn 19

Whitelaw and Argemon (2004) 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.

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 or personality.

As far back as 1753, Geroge-Louis Leclerc de Buffon declared that “Style is the man himself,” implying an inseparability between style from personality. John Middleton Murry wrote that “style is not an isolable quality of writing; it is writing itself” (1922).

Simpson (2004) 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.

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…” (1980). However, Walpole’s model of “style as option” does allow for both extra-linguistic analysis and even the possibility of style transformation.

Concerns with meaning creation, data interpretation and arbitrary meaning assignment form the heart of the major criticism leveled against stylistics (Carter 1988)(Simpson 2004). Stanley Fish (1981) calls stylistics’ procedures arbitrary and its conclusions either tautological or subjective. He considers the “establishment of a syntax personality” ultimately “meaningless” (Fiah 1981). Carter and Simpson in (1988) warn that stylistics analysis is too limiting since it has no access to the “extra-textual world of social, political, psychological, or historical forces.” Simpson relates a 1993 statement by linguist Jean-Jacques Lecercle who attacked stylistics as imprecise. “According to Lecercle nobody has ever really known what the term ‘stylistics’ means, and in any case, hardly anyone seems to care” (Simpson 2004).

4. STYLE AS OPTION

In “Style as Option” (1980), 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.'" (Walpole 1980)

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

Walpole also explains that style as "option" necessitates two assumptions. First, that the style is actually a conscious choice by the author and second, that it 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" (1980).

This view of style accommodates the every day 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 them 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. (1980)

She describes another exercise where students are encouraged to highlight and later subdue "obvious style" (1980).

5. STYLE MARKERS

Style markers, also referred to as "features", at their most basic, are indications of choices that were made in writing in a recognizable way. One objective of our project is to publish a comprehensive list of such markers.

Various works have offered examples of as well as process suggestions for working with style markers.

DeQuincey's notion of "mechanology of style," includes "counting sentence lengths, types and numbers of syntactic structures, and classes and varieties of words" (Walpole 1980).

Heidorn in (2000) 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. Almost all the options operate on the sentence level.

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 (Heidorn 2000) 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 (2005):

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

They also made the additional observation that "most studies are based on word forms and frequencies in occurrence".

An interesting example of feature #2 from above may be the Brill Tagger. Eric Brill (2000) 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 3 lexemes ahead and behind the current word. Rules are specified in terms of these variables and the current word-tag tuple being considered for alteration.

According to (Argamon 2003) "stylometric models have been based on hand-selected sets of content-independent, lexical, syntactic or complexity-based features."

In (Jucker 1992) 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." (Jucker 1992)

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. (Jucker 1992)

Latent Semantic Analysis (Landaur 1998) and Probabilistic Latent Semantic Analysis 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 natural language processing, computational linguistics and readability assessment. WordNet (2008), an online word ontology tool, can be used to calculate statistics of different word groupings. The Gunning fog index (Haardt 2007) and the Flesch-Kincaid readability test (Haardt 2007) are two popular means of assigning reading level by analyzing vocabulary in writing.

Authorship attribution work from Khosmood and Kurfess (2005); and Khosmood and Levinson (2006) contain an enumerated set of criteria used as features. In (Khosmood 2006) the authors found that vocabulary breadth was one of the best indicators of authorship. Their markers include vocabulary measurements (Fakotakis 2001), lexeme statistics (Kesselj 2003), as well as statistics based on grammatical constructs.

6. STYLE RECOGNITION AND TRANSFORMATION SYSTEM

The design for our system consists of three major components as shown in Figure 1. For classification, the system can store marker statistics for all analyzed corpora and it can find closest matches among those for a given source text.

For transformation, first the target corpus, typically made of many documents and the source text are both analyzed in terms of the presence of all or a subset of style markers. The system then performs a comparison between the source and target styles, calculates a stylistic distance. The Evaluator decides whether the comparison has yielded a match within pre-defined fuzzy boundaries. If not, the system chooses a transform consisting of one or more operators to apply to the source text in order to do modifications. Once the modifications are done, another comparison is made and if the styles are thought to be still too far apart, more transforms are chosen and applied until the system gets as closest possible to the source style, or if no more transforms are applicable in which case the latest version of the source becomes a "best effort" result of the transformation.

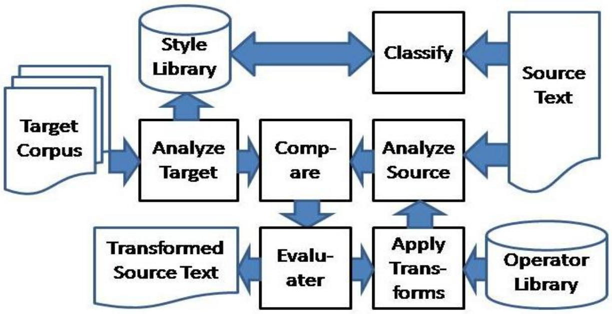


FIG. 1, overview of the system

7. ILLUSTRATIVE EXAMPLE

This is an example of what a style transformation system should be able to accomplish. The next section covers an experiment demonstrating the capabilities of our system.

Style S is 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 DeLillo and Toni Morrison should be enough to profile the style.

Function 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 formulation M1(t) > M1(s), where t and s are texts of style T and S respectively.

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

8. DEMONSTRATION

For this example, we chose as our target style, the first five chapters of George Orwell's Animal Farm which is 12426 words or 68,976 characters. The source text is an excerpt from a 2001 US Department of Justice memorandum Q&A text with questions removed*. It consists of 231 words or 1546 characters. The source text is the following:

> Second, agencies that have not already published recipient guidance should consider these factors and clarifications in preparing guidance documents. They should then submit their guidance documents to DOJ for approval prior to publication, as is required by the Executive Order. Following approval by the Department of Justice and before finalizing its guidance, each agency should obtain public comment on its proposed guidance documents. Those agencies also need to make the determinations regarding the Administrative Procedure Act and Executive Order 12866 as explained above.
>
> Third, as required by the Executive Order, agencies should continue to design and implement plans for making their own federally conducted programs and activities meaningfully accessible to LEP persons, and should consider the four-factor analysis from the DOJ guidance and today's memorandum in doing so.
>
> Federal financial assistance includes, but is not limited to, grants and loans of federal funds; grants or donations of federal property; training; details of federal personnel; or any agreement, arrangement, or other contract which has as one of its purposes the provision of assistance. If an agency does not engage in any of those activities, it does not grant federal financial assistance and does not have to issue a recipient guidance document. However, it must still design and implement a federally conducted plan to ensure access for LEP individuals to all of its federally conducted programs and activities (basically, everything that it does).

The system has 10 markers (Table 1). These markers were chosen mainly based on those textual measures necessary for the following readability indices: Kincaid formula, ARI, Coleman-Liau, Flesch reading easy formula, Fog index, Lix formula and SMOG grading level. The complete formulas and definitions for these indices are discussed in (Haardt 2007).

We could have chosen these indices themselves as style markers but we opted to use the most objective markers possible. Lix formula, for example is the average number of words in a typical sentence added to 100 times the ratio of words of three or more syllables to all words.

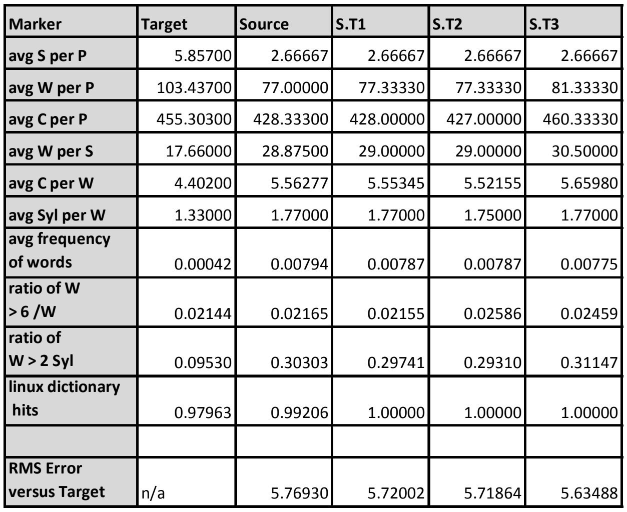
$$Lix = \frac{\#\, words}{\#\, sentences} + 100 * \frac{\#\, words\ with\ syllables > 2}{\#\, words}$$

FIG. 2, Lix formula

Thus while the Lix formula itself is not included as a marker, all of its building blocks are included in our set of markers. Another marker used is "Linux dictionary hit rate" which is a measured presence of words in the standard Linux dictionary. This measure is representative of vocabulary based markers. It could just as easily have been a hit rate of any other meaningful grouping of words. The Linux "look" utility is used to return all the entries in the dictionary, if any, corresponding to the word argument. If there is at least one entry returned from the dictionary, it is considered a "hit."

The system consists of three operators. T1 is a simple de-hyphenator operation. This operation simply removes any hyphen internal to the document usually creating two words where there was only one hyphenated word before. This allows a greater chance of each word being recognized by vocabulary based markers. It also changes words per sentence and words per paragraph statistics.

*http://www.usdoj.gov/crt/cor/lep/Oct26BackgroundQ&A.htm



Marker	Target	Source	S.T1	S.T2	S.T3
avg S per P	5.85700	2.66667	2.66667	2.66667	2.66667
avg W per P	103.43700	77.00000	77.33330	77.33330	81.33330
avg C per P	455.30300	428.33300	428.00000	427.00000	460.33330
avg W per S	17.66000	28.87500	29.00000	29.00000	30.50000
avg C per W	4.40200	5.56277	5.55345	5.52155	5.65980
avg Syl per W	1.33000	1.77000	1.77000	1.75000	1.77000
avg frequency of words	0.00042	0.00794	0.00787	0.00787	0.00775
ratio of W > 6 /W	0.02144	0.02165	0.02155	0.02586	0.02459
ratio of W > 2 Syl	0.09530	0.30303	0.29741	0.29310	0.31147
linux dictionary hits	0.97963	0.99206	1.00000	1.00000	1.00000
RMS Error versus Target	n/a	5.76930	5.72002	5.71864	5.63488

TABLE 1, matrix transformation of source through successive application of operators (S = sentence, W = word, P = paragraph, C = character, Syl = syllable, S.Tx = source after successive transformation x)

T2 applies a set of straight substitutions from GNU *diction* (Haardt 2007). *Diction* is a tool that tags various parts of text for possible clarity and rewrite reasons. It only makes straight substitution suggestions in a small number of cases. We identified 121 substitutions from the diction library and applied those to the text as part of T2. A few of the substitutions occur in our source text.

The T3 operator expands well-known acronyms such as DOJ (Department of Justice) and US (United States). As with all acronym expanders there are certain collisions possible. However, institutions such as major US agencies tend to consistently use one dictionary of acronyms with the additional stylistic requirement that the first use of such acronym should include the expansion. In this case, LEP (limited English proficiency) is a institution specific acronym which was defined earlier in the very document used as source text.

After application of each operator, the raw measurements are normalized based on the target corpus measurements, i.e. both sets of measurements are divided by the target ones. A root-mean-square error (RMSE) value is then calculated denoting the stylistic distance between source and target. Given two matrices S (source) and T (target), of n values each, our entire distance formula is the root-mean squared error (RMSE).

$$RMSE\ (normalized\ values) = \sqrt{\frac{\sum_i^n (1 - \frac{S_i}{T_i})^2}{n}}$$

FIG. 3, RMSE

In Table 1 the RMSE values are gradually reduced by a small amount after every application of a transform showing that the source style is moving slightly closer to that of the target, as detected via established markers.

Given more options in terms of styles and operators, the system continues with the same loop until more desirable results are obtained.

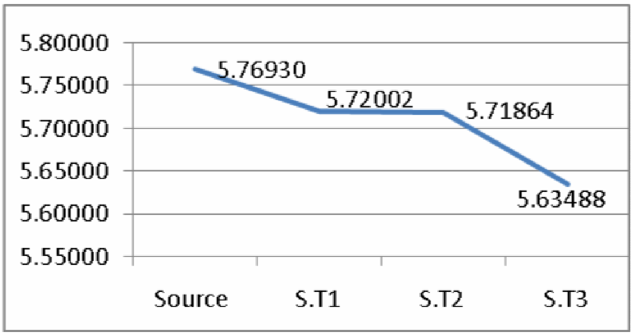


FIG. 4, declining RMS Error value (normalized data)

We now examine the annotated differences between S.T3 and Source texts emphasizing the transformations that have occurred within brackets [] below.

> Second, agencies that have not already published recipient guidance should consider these factors and clarifications in preparing guidance documents.
> They should then submit their guidance documents to [DOJ -> United States Department of Justice] for approval prior to publication, as is required by the Executive Order. Following approval by the Department of Justice and before ending its guidance, each agency should obtain public comment on its proposed guidance documents. Those agencies also need to make the determinations regarding the Administrative Procedure Act and Executive Order 12866 as explained above.
> Third, as required by the Executive Order, agencies should continue to design and implement plans for making their own federally conducted programs and activities [meaningfully -> meaningful] accessible to [LEP -> limited English proficiency] persons, and should consider the [four-factor -> four factor] analysis from the [DOJ -> United States Department of Justice] guidance and today's memoranda in doing so.
> Federal financial assistance includes, but is not limited to, grants and loans of federal funds; grants or donations of federal property; training; details of federal personnel; or any agreement, arrangement, or other contract which has as one of its purposes the provision of assistance. If an agency does not engage in any of those activities, it does not grant federal financial assistance and does not have to issue a recipient guidance document. However, it must still design and implement a federally conducted plan to assure access for [LEP -> limited English proficiency] to all of its federally conducted programs and activities (basically, everything that it does).

In one case above [meaningfully -> meaningful], the system does a transformation that is grammatically incorrect. This particular rule came from T2 based on the GNU *diction* suggestion list and it demonstrates potential pitfall with imperfect predefined rule sets.

9. CONCLUSIONS AND FUTURE WORK

We have examined related works in traditional linguistics, stylistics, literary studies and computational stylistics. We determined that one particular definition (style as option) is most suitable for computational systems.

Using the definition and ideas from the literature, we assembled an initial set of length-independent and machine-computable style makers which formed the basis of comparison between source and target styles. These were operationalized in our style classification/transformation system.

Our current results show that applying simple transformations to the source text can move the recognized source style slowly toward a target style, as measured by our definition of style. Future work will include more complex transformation rules applied intelligently and successively aiming for a greater move toward the target style.

REFERENCES

Argamon, S., Saric, M., and Stein, S., (2003) Style Mining of electronic messages for multiple authorship discrimination: first results. *Proceedings of the 9th ACM SIGKDD*, Washington DC.

Biber, Douglas (1989) A typology of English texts, *Linguistics* 27, 3–43.

Bradford, Richard (1997) *Stylistics*, Routlidge.

Brill, Eric (2000) Part-of-Speech Tagging, in *Handbook of Natural Language Processing* edited by Dale, Moisl and Somer, Marcel Dekker, Inc., pp 403-414.

Bringsjord, S., and Ferrucci (2000) *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Mahwah, NJ: Lawrence Erlbaum.

Comte de Buffon, (1773) *Discourse on Style*, trans. Rollo Walter Brown, in The Writer's Art, ed. *Brown, Harvard University Press*, 1921, pp. 285-86. (originally published 1773)

Carter, Ronald and Simpson, Paul (1988) *Language, Discourse and Literature: An Introductory Reader in Discourse Stylistics*, Routledge.

DiMarco, Chrysanne (1994) Stylistic Choice in Machine Translation. *Association of Machine Translation in the Americas Conference (AMTA)*.

Fakotakis, N. and Stamatatos, E. and Kokkinakis, G. (2001) Computer-based Attribution without Lexical Measures. *Computers and the Humanities*, Volume 35, Issue 2, May, pp. 193-214.

Ferrari, Giacomo, (2003) State of the art in Computational Linguistics. *Linguistics Today: Facing a greater Challenge*, International Congress of Linguists, John Benjamins Publishing Company, p 163.

Fish, Stanley (1981) 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.

Gervas, P. (2000), Wasp: Evaluation of different strategies for the automatic generation of Spanish verse. *Proceedings of the AISB00 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, Birmingham, UK.

Gonçalo, Hugo R. Oliveira; Cardoso, F. Amılcar and Pereir, Francisco C., (2007) Tra-la-Lyrics: An approach to generate text based on rhythm. *International Joint Workshop on Computer Creativity*, London.

Haardt, Michael, (2007) GNU *diction(1)* PDF manual, accompanying *diction* version 1.11. (http://www.gnu.org/software/diction/diction.html)

Heidorn, George E., Intelligent Writing Assistance, in *Handbook of Natural Language Processing* edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 181-209.

Karlgren, Jussi (2004) The wheres and whyfores for studying text genre computationally, *Style and Meaning in Language, Art, Music and Design*, Washington D.C. AAAI Symposium series.

Jucker, Andreas H. (1992) *Social Stylistics: syntactic variation in British newspapers*, Walter de Gruiyter.

Kesselj, Vlado et. al (2003) N gram-based Author Profiles for Authorship Attribution. *Proceedings of the Conference Pacific Association for Computational Linguistics*, PACLING'03, Dalhousie University, Halifax, Nova Scotia, Canada, August.

Khosmood, Foaad and Kurfess, Franz (2005) Automatic Source Attribution of Text: A Neural Networks Approach, *Proceedings of International Joint Conference on Neural Networks*, Montreal, Canada, June.

Khosmood, Foaad and Levinson, Robert (2006) Toward Unification of Source Attribution Processes and Techniques, Proceedings of IEEE International Conference on Machine Learning and Cybernetics, August.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998) Introduction to Latent Semantic Analysis, http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Loehr, Dan (1996) An Integration of a Pun Generator with a Natural Language Robot. *Proceedings of the International Workshop on Computational Humor*, Enschede, Netherlands. University of Twente.

Luyckx, Kim and Daelemans, Walter, (2005) Shallow text analysis and machine learning for authorship attribution. *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

Mairesse, Francois and Walker, Marilyn (2008) A personality-based Framework for Utterance Generation in Dialogue Applications. *Proceedings of the AAAI Spring Symposium on Emotion, Personality, and Social Behavior*, Palo Alto, March.

Moessner, Lilo (2001) Genre, Text Type, Style, Register, Terminological Maze? *European Journal of English Studies*, Vol. 5, No. 2, pp. 131–138.

Murry, John Middleton (1922) *The Problem of Style*. Oxford University Press, London, p. 77.

Rissanen, Matti et. al. (1991) *The Helsinki Corpus of English Texts*, (http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/)

Scigen (2005)- Scigen: an automatic CS paper generator. (http://pdos.csail.mit.edu)

Simpson, John, (2004) *Stylistics: A Resource Book for Students*, Routledge.

Walpole, Jane (1980) Style as Option, *College Composition and Communication*, vol. 31, No. 2, pp. 205-212.

Whitelaw, Casey and Argamon, Shlomo (2004) Systemic Functional Features in Stylistic Text Classification, AAAI *Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, October.

WordNet (2008) at Princeton University Cognitive Science Library, (http://wordnet.princeton.edu), accessed 9/2008.