Computational Approaches to Style and the Lexicon

by

Julian Brooke

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Computational Approaches to Style and the Lexicon

Julian Brooke
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2014

The role of the lexicon has been ignored or minimized in most work on computational stylistics. This research is an effort to fill that gap, demonstrating the key role that the lexicon plays in stylistic variation. In doing so, I bring together a number of diverse perspectives, including aesthetic, functional, and sociological aspects of style.

The first major contribution of the thesis is the creation of aesthetic stylistic lexical resources from large mixed-register corpora, adapting statistical techniques from approaches to topic and sentiment analysis. A key novelty of the work is that I consider multiple correlated styles in a single model. Next, I consider a variety of tasks that are relevant to style, in particular tasks relevant to genre and demographic variables, showing that the use of lexical resources compares well to more traditional approaches, in some cases offering information that is simply not available to a system based on surface features. Finally, I focus in on a single stylistic task, Native Language Identification (NLI), offering a novel method for deriving lexical information from native language texts, and using a cross-corpus supervised approach to show definitively that lexical features are key to high performance on this task.

Dedication

To Hong and Simon

Acknowledgements

First, I would like to thank my fellow students and post-docs in the Computational Linguistics Group at UofT for creating a community that has fostered this work: Libby Barak, Aditya Bhargava, Jackie Cheung, Paul Cook, Eric Corlett, Afsaneh Fazly, Vanessa Feng, Tim Fowler, Katie Fraser, Ulrich Germann, Siavash Kazemian, Varada Kolhatkar, Wesley May, Anthony McCallum, Aida Nematzadeh, Chris Parisien, Frank Rudzicz, and everyone else. A special thanks to my good friends Abdel-Rahman Mohamed and Tong Wang for some long discussions along the way (some of them even research related!). Outside the group, I'd also like to thank my collaborators Adam Hammond and Sali Tagliamonte for the opportunity to work with them and make this thesis truly interdisciplinary, and my external examiners, Jack Chambers and Moshe Koppel for their supportive feedback. Also thanks to Frasier Shein, Vivian Tsang, David Jacob, and all the other folks at Quillsoft Inc. for letting me join the team.

Thanks to my committee members, to Gerald Penn for putting up with my not-so-mathmatically-minded ways through four stimulating classes and as many checkpoints, and to Suzanne Stevenson for keeping me on task and then letting me know that I was actually done and should, yes, turn this thing in. To Graeme, of course, for more good advice on matters big and small than I can possibly even summarize here, and also for letting me ignore your advice now and again.

I also want to thank my Aunt Elizabeth, who contributed to this by being a lover of words and an unapologetic prescriptivist, and my parents, who actually made a few concrete contributions to the work in this thesis but mostly helped just by being available on Skype to hear me whine about it. Finally, thanks to Hong, for putting up with the weird hours and long trips, and to Simon, for really motivating me to get this done on time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

From a theoretical perspective, the underlying causes of stylistic variation in language, e.g. the mode, genre, writer/speaker background, intended audience, and aesthetic goals, are fairly distinct from the source of topic variation, which we take to be inextricably linked to the semantics, the message to be communicated. However, there is no easy line to be drawn in language itself, since both style and topic are accessed through the more fundamental layers of linguistic processing and production: phonology, morphology, syntax, and the lexicon. In computational linguistics, an artificial delimitation has been established, where the domain of topic (or semantics) is often considered to be the open-class content words, while stylistic analysis is mostly limited to "content independent features such as function words, part-of-speech and syntactic structures, and clause/sentence complexity measures" (Argamon and Koppel, 2010). In fact, in a survey of the field of computational stylistics, Argamon and Koppel (2010) explicitly claim that

> . . . textual features of style (as opposed to content) tend to function mostly in the aggregate—no single occurrence of a word or syntactic structure indicates style, but rather an aggregate preference for certain choices in a text rather than others.

A major goal of the research presented here is to move beyond this oversimplification to carry out a 'lexicalization' of style. Our methods—many of which are closely related to, though distinct in important ways from, those originally developed for modeling topics—will be used to acquire information about the style of individual lexical items, and show that this information both corresponds to human intuitions and is useful when applied to relevant tasks in computational stylistics. In doing so, I demonstrate that the lexicon is in fact fundamental to stylistic variation.

In this thesis, I focus on three perspectives on style. The first, and perhaps most common, views style as being related to aesthetic choices of a writer; this is the approach taken by, for instance, prescriptivists interested in improving student writing quality. A more descriptive view is offered by linguists who study genre and register, where style is a reflection of the functional requirements of various text types. A third view considers the demographics or biographical details of the author, e.g. age or native language, as being of paramount importance; this bring us into the territory of sociolinguistics and second language acquisition. Importantly, the lexicon has already played a major role in the first of these perspectives, but is underdeveloped or ignored in the other two. Here, I will show that the lexicon serves as an important link between these different conceptions of style; by drawing these connections, we can ground vague layperson intuitions about aesthetic style in tasks which reflect real-world variation, and at the same time go beyond a narrow focus on individual stylistic tasks to identify generalizations that are not just useful for improving performance on these tasks, but that can also be of benefit to researchers working in related fields such as linguistics, literary analysis, and education.

Another of the tenets of this thesis, then, is that there is value in looking at style as a broader phenomenon, rather than as simply a motley collection of tasks. Although there is no one-size-fits-all model for the entire space—indeed, I will be applying a wide range of statistical and vector-space approaches—there are themes that will reappear regularly throughout this thesis,

and connections can be made that justify the more holistic vision being offered. One example of this is the area of feature representation: traditional stylistic work as well as work in topic modeling has tended towards some type of normalized frequency-based features, but I will show that this is not appropriate for lexicalized style. Another key theme of this work is a distrust of certain kinds of traditional evaluation, particularly within-corpus evaluation (e.g. cross-validation, perplexity measures); stylistic phenomena are inherently high-level, but all corpora, as a result of the specific methodology of their construction or collection, have other broad patterns which are not stylistic. If our focus is too myopic, limited to optimization of a single task in a single corpus, we cannot be certain our models have really grasped anything about style at all. Here, my focus is explicitly broader, and as a result I will prefer, when possible, human interpretability, unsupervised methods in large, diverse corpora, and evaluations across multiple tasks and multiple corpora.

The work in this thesis will address a wide range of stylistic phenomena, but our definition is not an all-encompassing one. First, our lexical focus will limit the discussion to some degree, even though non-lexical features will be considered at various points throughout. Second, and perhaps more importantly, we are not interested here in an overly personalized definition of style, such as one might assume in the context of authorship attribution (Stamatatos, 2009b). Although individuals do of course have particular lexical preferences that could be called their 'style', in rare cases being human 'interpretable' in the sense of being distinct enough to allow for identification, these features are clearly not appropriate targets for lexicalization as intended here; we are looking to capture much more general linguistic regularities.

In Chapter 2, I review work in linguistics relevant to style, including prescriptive manuals of style, both theoretical and applied work in register and genre, and sociolinguistics.

Chapter 3 is concerned with the induction of stylistic lexicons from corpora. First, I summarize work in two related areas, topic modeling and polarity lexicon induction. Then, I present

my own work using these methods to build lexicons for formality, readability, and sociolinguistic variables (e.g. age, gender). The centerpiece of this section is a method for concurrent induction of multiple correlated styles.

In Chapter 4, I go beyond lexical items to discuss a range of stylistic tasks which can use the kinds of resources built in Chapter 3. The first section is a review of previous work relevant to various tasks, not all of which I will address directly myself. My own work in this area includes a word choice task, genre differentiation, and within-text stylistic inconsistency segmentation and clustering, particularly as applied to literature.

Native language identification could arguably be included under the stylistic tasks of Chapter 4, but it differs from those considerably in that human-interpretable stylistic dimensions are not the key source of variation. Due to problems with the popular corpus for this task, lexical (n-gram) features for supervised classification had been overlooked until recently; in Chapter 5, I use alternative, cross-corpus evaluation to show that lexical features are key to this stylistic task, and that other kinds of stylistic variation are important in this space.

Chapter 2

Linguistic Foundations of Style

Perhaps due to its associations beyond language, the term *style* is most often used to refer to the aesthetics of a written text, most often created by the conscious efforts of a skilled writer or the unconscious failure of a poor one; this conception of style is clearly visible within the field of prescriptive linguistics, which we discuss in the first section below. However, style has deeper roots in linguistics proper (i.e. descriptive linguistics): the true underpinnings of style, in terms of the differences in co-occurrence that allow for these aesthetic judgments, are variations across text types, i.e. register and genre, and the variations across groups of individuals, i.e. sociolinguistics.

2.1 Prescriptive stylistics

Near the end of the 3rd edition of the classic style handbook, *The Elements of Style* (Strunk and White, 1979), E.B. White defends their prescriptive approach to writing education:

> The intent is to suggest that in choosing between the formal and the informal, the regular and the offbeat, the general and the special, the orthodox and the heretical, the beginner err on the side of conservatism, on the side of established usage. No

> idiom is taboo, no accent forbidden; there is simply a better chance of doing well if the writer holds a steady course, enters the stream of English quietly, and does not thrash about.

Style and usage manuals are well known for categorizing and codifying the details of written language, providing collections of idiosyncratic edicts that are often at odds with common usage even at the time, sliding into pure anachronism sometime thereafter, see Pullum (2009) for a modern critique of *The Elements of Style*, and Milroy and Milroy (1999) for a critical look at linguistic prescriptivism in general. In this respect, other popular usage manuals (Fowler, 1968; Follett, 1966) are more problematic than *The Elements of Style* (hereafter *Elements*), because they are much more encyclopedic in their detail, and thus even more susceptible to individual eccentricities. Nevertheless, such books offer a useful starting point for the present work, since they are unabashedly concerned with written style. For instance, the above quotation suggests a straightforward definition, one that will be reinforced throughout this section: rather than being a set of stodgy rules, style is characterized as a series of lexical and grammatical choices that have varying pragmatic consequences. A good text, then, is one where the writer makes appropriate and consistent choices at each step of the process (Follett, 1966, I.I). In the above quote, White is essentially arguing that the best option for a novice writer is to follow the generally accepted practice whenever possible; it is better to make a 'standard' choice than a risky one. Using examples from these manuals, I will highlight here some of the key stylistic decisions that writers face, collecting some initial judgments of what might be considered style.

Perhaps foremost among the concerns of a modern style manual is making the text easily understood to the reader: "Be clear" (V.15, *Elements*), "Do not take shortcuts at the cost of clarity" (V.19, *Elements*). The rule "Keep related words together" (II.20, *Elements*) is intended to help the writer avoid ambiguity, as are "Make sure the reader knows who is speaking" (V.13, *Elements*) and the call for writers to boycott the (now) common senses of certain words,

e.g. *hopefully*, *facility*, *presently*, and *transpire* (IV, *Elements*). In another well-known style manual, *The King's English* (*King's*) (Fowler and Fowler, 1906), the first rule for lexical choice is "prefer the familiar word to the far-fetched" (Chapter I): there can be no clarity if you are using vocabulary your reader does not know. In a third (Follett, 1966), the cardinal principle of good writing is that "no one should ever have to read a sentence twice because of the way it is put together" (III.2). However, there is no universal consensus among style mavens on this point; for instance, Lanham (1974) argues that clear language is quite often dull, ugly language, and there is a clear place for obscurity in a writer's repertoire.

Two related stylistic virtues are conciseness and lack of pretension: The examples for "Omit unnecessarily words" (II.17, *Elements*) include the substitution of simple expressions for complex, wordy ones, e.g. "since" instead of "owing to the fact that" and "remind you" instead of "call your attention to the fact that." In section V, the student is urged to avoid "rich, ornate prose," and "fancy words". Similarly, the third and fourth rules of *King's* are "prefer the single word to the circumlocution" and "prefer the short word to the long" (chapter I). In *The Oxford Guide to Writing* (*Oxford*) (Kane, 1983), Kane devotes a chapter to unnecessary words, and suggests that unusual words (including foreign borrowings) be "used with caution," to avoid appearing pretentious (Chapter 42).

Using complex wording and excessive verbiage might result in a formal, stilted text, but *Elements* also warns strongly against the opposite extreme: "Do not affect a breezy manner" (V.9). Formality is a consideration throughout *Elements*, for instance the dash is introduced as "less formal than a colon, and more relaxed than parentheses," and several words and expressions are dismissed as being too colloquial (e.g. *fix*, in the sense of *repair*, and *sort of* and *kind of* as adverbial modifiers). In *Oxford*, students are warned to be wary of colloquialisms, which are often vague, and the mixing of informal and formal language "must be exercised responsibly," though the skilled author can use this to "striking" or "comic" effect (Chapter

44). Oversimplification is also a possible misstep: in *Elements*, the student is warned to "avoid a succession of loose sentences," (II.18) i.e. sentences with two clauses, one subordinate to the other. Several chapters of *Oxford* are focused on style of sentence construction, including sentence variety and rhythm (Chapters 38, 39). In the same vein, *King's* notes that some variation in the vocabulary used to describe some referent (known as (in)elegant variation) is desirable and even necessary to avoid dull repetition, but too much can seem both pretentious and make the text difficult to follow (Chapter III).

In the various guides, the student is urged to "use definite, specific, concrete language" (II.16, *Elements*), to "prefer the concrete word to the abstract" (I.2, *King's*), and to "make your words as concrete and specific as the topic allows" (Chapter 41, *Oxford*). In *Elements*, positive forms are encouraged, since the negative often results in an indefinite, hedging expression: the positive "distrusted" as compared to the negated "did not have much confidence in"; likewise, passive expressions often result in indefiniteness and ambiguity, thus lacking authority. Several words are singled as being particularly vague, for instance *contact*, *nice*, and *offputting*. In *Oxford*, students are told to avoid generic words that indicate a class of things, preferring, when possible, specific words like *terror* to hypernyms *emotion* or *fear*.

Another aspect of style relates to the influence of the writer's personal biases: "Do not inject opinion" (V.14, *Elements*); "Place yourself in the background" (V.1 *Elements*). Student writers should maintain objectivity with respect to the subject, being careful that subjective emotions do not color their language. In his critique of *Elements*, Pullum dismisses this rule as "truly silly" (Pullum, 2009). Indeed, the rule, although appropriate for some genres of writing (e.g. a newspaper article), seems entirely unsuited to others (e.g. a personal narrative). That said, there is a clear relationship between objectivity and authority in society, and this authority is undermined, for instance, when a writer engages in exaggeration: "Do not overstate" (V.7, *Elements*); "false hyperboles. . . rather than impressing us with the importance of the sub-

ject, . . . make us laugh at it"' (Chapter 41, *Oxford*).[1] An example of hyperbole from *Oxford*: "Football is the most magnificent sport ever developed by the mind of man."

In the discussion of "Write in a way that comes naturally" (V.14, *Elements*), White notes that all language relies on subtle imitation; his goal here is to encourage writers to avoid obvious mimicry, and to instead to cultivate a natural style by exposure to good writing. One result of limited experience in the (written) language is what Kane calls "wrong idiom," which includes problems with prepositions, fixed verb/object combinations, and other collocations, e.g. "we have a *great* (idiomatic: *high*) standard of living" (*Oxford*, Ch 41). For non-native writers, natural patterns of usage from their first language may be carried over into their new language (Odlin, 1989); not only can this affect clarity, but it may have other undesirable pragmatic effects in the mind of the reader, in the extreme it may lead to discrimination (Milroy and Milroy, 1999; Campbell and Roberts, 2007). Non-native speakers, often armed with a limited linguistic arsenal, are also susceptible to another stylistic error: the use of clichéd or hackneyed language, e.g. "white as snow" (*Oxford*, Ch 41), "the foreseeable future" (*Elements*, Ch 41).

The discussion here has far from exhausted the advice contained in these and other popular writing guides (Gunning, 1952; Williams, 1990; University of Chicago, 2003; Garner, 2009); here I have explicitly ignored the fine details of mechanics that are often highlighted (e.g. *which* vs. *that*, or *will* vs. *shall*) as well as extremely general properties of writing that would not be directly observable in lexical or grammatical choice (e.g. textual organization). From the perspective of computational applications, our conception of style must have a clear empirical basis, and should apply to language in a broader, more theoretically satisfying way than is possible from within the prescriptivist framework. Thus we move next to work on style within

[1] It is worth mentioning that languages may vary in type, amount, and intensity of emotion that is generally expressed in a particular context: see the cross-cultural observations in Bautin et al. (2008). It follows that non-native speakers who are relying on the standards of their first language may create texts that seem exaggerated (or understated) to native speakers; in fact I have personally witnessed this phenomenon in the essays of Chinese EFL learners.

the (descriptive) linguistics literature, where context plays a fundamental role.

2.2 Genre and register

The work of Joos (1961) represents an initial attempt to systematize the notion of style outside of a normative context. He proposes basic five styles of English: 'Frozen' (unchanging printed language); 'Formal' (Non-interactive, one-way interaction); 'Consultive' (interactive, cooperative); 'Casual' (between friends and acquaintances); and 'Intimate' (private among intimates). In contrast to the edicts of the prescriptive approach, Joos notes that each of these styles has an appropriate context.

Crystal and Davy (1969) apply linguistic analysis directly to the study of style in non-literary texts, both written and spoken. The major focus is on the "dimensions of situational constraint", namely the extra-linguistic facts of the discourse. The latter consists of relatively immutable social factors, such as regional/national origin, class, gender, and individual style of the speaker/writer(s), as well as discourse-specific features such as time, medium, topic ('province'), genre ('modality'), and the social relationships involved ('status'). Each dimension has a number of categories that can be associated with it: for instance, *formal* or *informal* is included under the *status* label. One interesting aspect of the work is the focus on the interdependence of various categories, for instance the mutual dependence of 'legal' and 'formal', the probable co-occurrence of 'conversational' and 'informal' and the highly improbable co-occurrence of 'legal' and 'colloquial'; these interdependencies require a detailed analysis of multiple texts in multiple genres, in order to differentiate the relevant contexts of features that otherwise co-occur.

Crystal and Davy's particular definition of 'situational constraints' is only one of many possible formulations of contexts. Probably the most widely known is the triple of *Field*,

Tenor, and *Mode*, which stems from work by Gregory and others (Gregory and Carroll, 1978) but was adopted as part of *Systemic Functional Linguistics* (SFL) (Halliday, 1994). Under the original definition, the term *field* encapsulates information about the institutional setting and subject matter of the communication, *tenor* includes information about the participants in the communication, and *mode* is the function of the text relevant to the communication event including its channel (e.g. written/spoken) as well as its genre (form). As noted in the critique by van Dijk (2008, Chapter 2), these are rather vague definitions, collapsing some of Crystal and Davy's dimensions into a single category, seemingly for convenience of notation and to preserve analogy with SFL's three functions of language: the *ideational*, *interpersonal*, and *textual*. Under this theory of context, a particular configuration of tenor, field, and mode is associated with a set of corresponding linguistic features, defining a *register*, and a coherent text must be consistent with respect to its register (Halliday and Hasan, 1976).

Leckie-Tarry (1995) develops perhaps the most detailed theory of register within the SFL framework, using the triple of field/tenor/mode as the starting point for a conception of context with significantly more depth. Fundamental to the theory is a (new) three-way distinction between the context of situation, the context of the text (co-text), and the context of culture. The central (continuous) dimension or *cline* of register is more or less defined by full reliance on the context of situation at one end, the oral pole, and full reliance on the context of culture on the other, the literate pole. Elements of context (essentially a decomposition of field/tenor/mode) and the linguistic features that realize them are also represented as clines, running 'parallel' to the main cline of register, partially dependent on it and each other, "establishing probabilistic relationships between a given register and certain lexical, syntactic, and discourse structures which may be used to realize it" (Birch, 1995). This is somewhat more theoretically satisfying than the haphazard connections between variables in Crystal and Davy (1969); it is important to note, however, that Leckie-Tarry does not provide an explicit mathematical model or empirical

results to ground her conception of register.

It is illuminating to compare Leckie-Tarry's theory, for instance, with the approach of Paolillo (2000), who, working within the Head-driven Phrase Structure Grammar framework, adds discrete 'communicative attitudes' to the HPSG feature structures that condition particular syntax choices (i.e. make them 'unifiable' or not, in the given discourse context) with the goal of capturing register variation in the Sinhala language of Sri Lanka under a single grammar (previous approaches had treated Sinhala as being strictly diglossic in character). While offering a formal basis that is lacking in Leckie-Tarry's work, Paolillo's approach is less theoretically satisfying in the sense that the 'communicative attitudes' do not, as one would hope, vary independently to define all possible registers: rather, they apparently switch sequentially as one 'moves' from more 'literary' to 'conversational' registers; in essence, they define a step-wise version of Leckie-Tarry's continuous range, including cases where relevant syntactic features are found in both forms in the same text.

The contextual clines in Leckie-Tarry's theory include variables that seem appropriate to a continuous paradigm, for instance degree of specialization, power, planning, (physical) distance, and the education, class, age, and intelligence of the participants, though not all contextual variables seem intuitively scalable (*medium*, for instance, seems inherently discrete, i.e. either written or spoken, though Leckie-Terry provides a cline nonetheless). As for the clines of linguistic realization, they involve some abstraction above simple linguistic features, for instance: the cline of 'Generalization,' which varies from verbal/clausal at the oral pole to nominal/lexical at the written pole; the cline of 'Syntactization,' where the oral pole is characterized by attention to topic/focus (i.e. pragmatic concerns) while the structure at the written pole is determined by semantic role relations; the cline of 'taxis,' with coordinated syntax at the oral pole and subordinated syntax at the written; the cline of 'lexis,' with informal words at the oral end, formal words at the written end, and the core vocabulary (Carter, 1998) in the

middle; and the cline of 'Information,' from specific dynamic at the oral pole to generic stative at the written.

One key complaint raised by van Dijk (2008) against the various SFL conceptions of context is that, by taking a functional stance, SFL theory draws too strong a connection between facts in the world and their realizations in language; in van Dijk's view, it is not the situation itself that directly determines the stylistic choices of a speaker/writer, but rather his or her *mental model* of the situation as well as his or her ability to translate that model into the 'appropriate' linguistic output. Mental models (Johnson-Laird, 1983) are intended to be comprehensive representations of the relevant context, going well beyond the social context to include various kinds of world knowledge that aid in constructing and interpreting discourse. Certain aspects of such a model are similar to the other contextual models we have seen here; however, adding a layer of subjectivity allows for a clash between the mental models of the participants. For example, I might choose a certain style that, in my model of your model (mental models naturally contain approximations of other mental models), would have the effect of increasing your estimation of my level of intelligence, but the actual effect in your model is the opposite. Though very appealing theoretically, the power of such a model becomes a liability if we consider it in the context of computational applications.

On the other extreme of the spectrum, Multi-Dimensional (MD) analysis (Biber, 1988; Biber, 1995; Biber, 2006) makes no prior assumptions about the relevant contextual variables; instead, the most important dimensions are derived from the data using factor analysis on a mixed-register corpus. Given sets of features that have positive or negative loadings for each dimension, it is usually possible to provide a qualitative description of the dimension. After an analysis of the Lancaster-Oslo-Bergen corpus using features derived primarily from Quirk et al. (1985), Biber (1988) identified six dimensions of variation in English: *Involved vs. Informational*, characterized by the presence or absence of verbs and first- and second-person

pronouns and fewer nouns and long words; *Narrative vs. Non-narrative*, characterized by past-tense verbs and third-person pronouns; *Situation-dependent reference vs. Context-independent reference*, characterized by time and place adverbials and a lack of complex noun phrases; *Argumentative vs. Non-Argumentative*, characterized by infinitives, modals, and conditional subordination; *Abstract vs. Non-Abstract*, characterized by conjuncts, passives, and past-participial clauses; and *On-line informational elaboration*, characterized by *that* clauses. Texts of varying register (genre) were ranked for each dimension; telephone conversations, for instance, were found to be the most 'Involved,' and official documents the most 'Informational.' MD analyses of (English-language) university texts (Biber, 2006) and diverse texts in other languages (Biber, 1995) found a related but distinct set of dimensions, suggesting the MD paradigm is fairly corpus dependent, and might not be the best way to identify 'universal' dimensions of register; in the case of other languages, however, it could also reflect particular cultural concerns, for instance a dimension in Korean which seems to closely linked to politeness.

2.3 Sociolinguistics

Variationalist sociolinguistics (Labov, 1972; Trudgill, 2000; Tagliamonte, 2011) is another empirically grounded approach to exploring the influence of contextual features on style. Originally concerned with phonological differences in American dialects, the field has expanded to include a wide range of sociological and linguistic variables. Though differences between any distinct social groups which speak the same language are potentially of interest, three of the most social groupings are gender, class, and age; the last, in particular, can be used to trace the change of the language over time.

The methodology of variationalists generally includes field work to collect data (speech) from linguistic communities that vary in one or more contextual factors that are thought to

influence a particular linguistic variable. Sociolinguistics research traditionally prefers spontaneous spoken language, though historical research is of course dependent on written records, and recent work has expanded to include new genres like internet chat. A variable is defined as a set of linguistic alternatives that do not affect the underlying semantics of what is being expressed, e.g. the choice of *gotta* versus *got to*. All instances of each variable in the data are identified, and the predictive factors (both social and linguistic) for each are enumerated. The final step is to build a logistic regression model over these instances; the goal is not to predict new instances of the variable, but rather to see which factors contribute to the decision more than would be predicted by chance, i.e. which are statistically significant. This information is often used to come to sociological conclusions about change in the linguistic community.

One interesting distinction made among variables (or rather, forms of variables) in sociolinguistics is how amenable they are to conscious examination or change. *Indicators* are forms that are outside of the control of speakers; they are used even in cases when a speaker is obviously trying to avoid using group-specific language. By contrast, *Markers* are often modulated by the demands of the situation, though a speaker may be unaware of the specific choices he or she is making to create stylistic effects (e.g. more or less formal). On the other hand, *Stereotypes* are forms that are universally recognized as being distinctive to a particular social group, and as such may be embraced or avoided depending on their prestige. Within a speech community, there is a tendency for forms to go from indicator to marker to stereotype (Labov, 1972). Indicators and markers are, by definition, subtle (they haven't been fully recognized as distinctive); an example I come back to later in my own work in Chapter 3 is the word *supper* (rather than *dinner*) as an indicator of age. By contrast, Canadian raising (Chambers, 2006) and the use of discourse *be like* to introduce speech among teenage girls (Tagliamonte, 2005) have both obviously progressed to the stereotype stage. This distinction is important, since one major interest of my work here is in building human-interpretable lexicons; however, 'style'

in computational linguistics has often been focused on aspects of language (e.g. frequency of function words) which are really only accessible via statistical analysis, corresponding to indicators rather than markers or stereotypes.

Sociolinguistics research is typically focused on individual variables, many of which are not particularly lexical. In the context of our lexical interest here, a full review of trends in sociolinguistics research would be a detour, but it is worthwhile to look at one relevant variable that has received recent interest: *intensifiers*. Relevant research (Ito and Tagliamonte, 2003; Tagliamonte, 2011) suggests that the intensifier *very* is quickly falling out of use among the young (and thereby picking up a stodgy feel), *really* is now the preferred intensifier in North America, and *so* and *pretty* (as intensifiers) are less common but are gaining traction among young women and young men, respectively—the fact that *so* is much more forceful than *pretty* reflects differences in social expectations of gender (i.e. showing strong emotion versus maintaining emotional distance). In this case, it seems clear to me that social dimensions such as gender and age can be more or less directly linked to stylistic dimensions such as subjectivity and formality.

Chapter 3

Stylistic Lexicon Induction

This section deals with the identification and/or quantification of stylistic aspects of the lexicon using automated methods. First, we motivate this need by discussing style in the context of existing lexical resources, and then turn to the computational modeling of topic and polarity that is relevant to stylistic modeling. The rest of this section presents new work in stylistic lexicon induction. We summarize these contributions in Table 3.1.

3.1 Style in the lexicon

Automated identification of the semantics of words, for instance various *-onymy* relationships among words that are manually annotated in a resource such as WordNet (Fellbaum, 1998), is a fairly well-addressed (though far from solved) problem in computational linguistics (Hearst, 1998); by contrast, the stylistic facet of the lexicon is comparatively unexplored. This is not a shocking oversight, of course, since the substantive meaning of a text *is* generally more important than the mere 'way it is put'. And, if so, why not focus our collective attention on solving semantics first? One response is that discovery of stylistic variation might actually be more amenable to the kinds of statistical models that have come to dominate the field. In

Table 3.1: Overview of contributions in Chapter 3, including aspects of styles investigated, methods used, and conclusions reached.

- Section 3.4

  Aspect Formality

  Methods Frequency, corpus ratios, PMI, LSA, hybrid combinations

  Conclusions LSA best individual method, binary features preferred, lower dimensional vector 'stylistic', improvement from combining, different sources of information, overall results promising

- Section 3.5

  Aspect Readability

  Methods Frequency, corpus ratios, average document metrics, crowdsourced evaluation

  Conclusions All metrics useful, major boost from combination, LSA and document metrics redundant, automated agreement near human agreement

- Section 3.6

  Aspects Literary,colloquial, abstract, concrete, objective, subjective

  Methods LDA, correlated topic model (CTM) inference in BNC, correlation analysis

  Conclusions Binary features preferred, multi-style model better, strong correlations among styles as predicted by polar model, LDA better than CTM, objective and abstract are difficult to distinguish

- Section 3.7

  Aspects Literary, colloquial, abstract, concrete, objective, subjective

  Methods Annotation, Kappa, LDA, LSA, NPMI, label propagation, linear regression

  Conclusions Disagreement as indicator of scale, LSA better for majority of styles, LDA better for most common, scores can be further refined with supervised methods

- Section 3.8

  Aspects Age, education, work

  Methods Feature selection, mutual information

  Conclusions Corresponds to known patterns, some interesting new variables, manual effort required

particular, the social variables underlying stylistic variation are much more tractable than those underlying semantic variation: the latter ultimately reflects the full complexity of the external world which language presumably evolved to describe, while the former is grounded in a social world which can be conceived of as a modest set of key traits or dimensions, such as social role and social distance. Moreover, a better grasp of stylistic variation may lead to improvement in

our models of semantics; at the very least, it may be possible to eliminate this stylistic 'noise' from such models. Given that the most convenient source of large-scale linguistic data, the World Wide Web, contains variation that reflects nearly the full social range of the language, a better understanding of this variation seems long overdue.

The branch of descriptive linguistics concerned with style and register, which I have discussed in some detail earlier in this work (Section 2.2), is not overly concerned with specific lexical items. The feature set of Biber (1988), which has influenced much work in computational linguistics, considers content words mostly in the abstract, counting the frequency of particular parts of speech or the number of words longer than than a certain cutoff (which are supposed to be technical terms). Similarly, the 'contextuality' (formality) metric of Heylighen and Dewaele (2002) is POS-based. Other important work in this area, for instance that of Leckie-Tarry (1995) is primarily focused on building a theoretical framework for understanding the influence of context; details of realization are not of interest. Sociolinguistics, the study of language variation and change, provides an alternative framework for understanding stylistic differences (see discussion in Section 2.3); although individual linguistic features play a key role in the analysis, the field has traditionally been focused on phonological and syntactic variation, since one of the key requirements for a 'variable' as defined by researchers in the field is that the alternatives be semantically interchangeable, which is rarely true for lexical items. Studies of readability are also focused primarily on simple textual metrics (van Oosten et al., 2010), though there are a few recent exceptions that use age-tagged texts to derive the grade levels of individual words (Kidwell et al., 2009), or identify core vocabulary (Li and Feng, 2011).

By contrast, many works of English prescriptive linguistics, e.g. *The Elements of Style* (Strunk and White, 1979), address the style of lexical items, at various levels of specificity. At the general level, there is the divide between the everyday Anglo-Saxon vocabulary, which

forms the core of the English language, and the Romance (mostly Latin and French) vocabulary that for historical reasons became specialized to the intellectual and artistic life of the educated elite (Fowler and Fowler, 1906; Williams, 1990). Williams, for instance, describes the former as clear, direct, concrete, readable, and plainspoken, while the latter is turgid, flamboyant, complex, abstract, unreadable, passive, and pretentious. Yet bad writers (and orators) cannot, apparently, resist overusing Romance vocabulary, in the hopes of appearing educated. Additionally, it is worth noting that some Romance vocabulary is so integrated into the core English language that it no longer carries any particular stylistic connotation; it is a rough distinction, not a categorical one, and one that provides little or no assistance in other languages.[1]

Specific words and multi-word expressions are singled out for special attention in many usage manuals; for some this is clearly the focus of the book (Fowler, 1968; Follett, 1966). Often, these are warn writers to avoid misusing or abusing terms. One particular usage manual that has received attention from us (see Section 3.4) and others is *Choose the Right Word* (Hayakawa, 1994), which provides a comparison of various near-synonyms, highlighting connotational and denotational differences. The stylistic information included in entries is necessarily relative, rather than absolute, and the resource is generally limited to those lexemes which have numerous near-synonyms, so it is far from comprehensive. Dictionaries often have tags that correspond to stylistic attributes (e.g. vulgar, slang, archaic, literary), but these usually apply only to the most extreme instances, and these dictionaries usually include only a limited set of multi-word expressions. Slang and idiom dictionaries such as the *Urban Dictionary* often provide an in-depth look at this particular stylistic dimension, but there are important stylistic distinctions within this general category, for instance how vulgar or up-to-date the slang term is. Other sizable lists of emotionally-charged language are included in text analysis resources such as the General Inquirer (Stone et al., 1966) and LIWC software[2], and there has been

[1] Though obviously many Latinate terms have similar stylistic functions in other European languages.

[2] http://www.liwc.net

much work on automatically or semi-automatically generating sentiment/emotion lexicons for use in sentiment analysis (Taboada et al., 2011). Some elements of style, e.g. familiarity and concreteness, are captured for a small (1000+ word) vocabulary in the MRC psycholinguistic lexicon (Coltheart, 1980). However, the most popular (and comprehensive) lexical resource in computational linguistics, WordNet (Fellbaum, 1998), does not capture stylistic variation at all, since stylistic differences occur within the *synset* unit.

3.2 Models of topic

Perhaps the most straightforward approach to topic modeling is supervised text classification using a simple bag-of-words feature set, where the topic of a text is the output of the trained classifier. Indeed, topic classification of this kind, within the relatively limited domain of Reuters newspaper articles, was an early benchmark for text classification as a field (Sebastiani, 2002). The most active area of research within text classification, at least initially, was the introduction and comparison of different classification algorithms, including decision trees, naive Bayes, support vector machines, maximum-entropy, neural networks, example-based classification, and classifier committees (e.g. AdaBoost). Dumais et al. (1998), for example, compare several popular options and conclude that SVMs are the most promising for topic classification. Though supervised algorithms will play a role in this research, lexical acquisition, in the context of a machine learning framework, is usually considered a facet of feature selection or dimensionality reduction, consisting primarily of unsupervised or weakly semi-supervised methods.

Certain simple but extremely popular methods for feature manipulation warrant some discussion here, since we will be re-evaluating them in the light of our stylistic (rather than topical) goals. Typically, one of the first steps in topic classification or information retrieval is the removal of extremely common words, i.e. a stop list, presumably because they carry no important

topical information. This contrasts sharply with work in computational stylistics, where function words are often primary features and sometimes the only features in the analysis (Koppel and Ordan, 2011). At the other end of the spectrum, very rare words are often excluded since they explode the size of the vocabulary (due to its Zipfian distribution) and yet provide little information, particularly in the context of machine learning. One extremely popular measure that combines these two approaches is *tf·idf* (Jones, 1972). Defined for a term within a text that is itself in part a larger text collection, the weight of a term under *tf·idf* is defined to be its how often it appears in a given document (words that are appear often in a text are important to the text) divided by the number of other documents it appears in (if a word appears everywhere, it is not that important). Though not explicitly a metric of topic, one effect of *tf·idf* is that it will weight words strongly associated with the topic of a text, and deemphasize words that are likely not (including typical stop words); it has, for instance, been used to remove topical elements in cases where they might be a confounding factor (Tsur and Rappoport, 2007). There are several variants of *tf·idf*, including logarithmic calculations of the frequencies, and other kinds of normalization (Manning et al., 2008).

Term weighting measures such as *tf·idf* play a key role in the vector space model (Salton et al., 1975), most commonly associated with information retrieval. In this paradigm, documents (or queries) are represented as a vector of dimensionality equal to the size of the vocabulary (term-vectors), with weights as their values. Documents can be then compared using various metrics, for instance the cosine of the angle between their vectors, or the Euclidean distance. If the values (the weights) are strongly sensitive to topic, the vector space can thus be considered a model of topic, documents of similar topic will generally be close to each other in this space. It is worth noting that strong stylistic variation may interfere with the vector space model's ability to model 'true' topic, if the difference results in entirely different set of vocabulary being used: for instance, a clinical discussion of drug addiction by a medical expert may have

very little (topical) vocabulary overlap with an addict talking to a fellow addict about his need.

Though in the context of information retrieval, the vector space model is mostly limited to document vectors, in the broader context of computational linguistics the idea has been turned on its head with the study of vector space semantics. There, the focus is on the representation of the terms (words), defined in terms of the documents they appear in, or other features (terms, patterns) which appear in the same document, paragraph, sentence (Turney and Pantel, 2010). Though still typically *bag-of-words*, some approaches integrate syntactic relationships (patterns) (Lin and Pantel, 2001), and offer some treatment of compositionality (Erk and Padó, 2008). These vectors can be used for various applications, in particular measuring word similarity (Landauer and Dumais, 1997) and word sense disambiguation (Yuret and Yatbaz, 2010).

Vector space methods often rely crucially on some form of dimensionality reduction (or matrix smoothing), that is, a reduction in the complexity of the space. Beyond the more efficient processing offered by the shorter vectors, one goal of dimensionality reduction is identifying the most important latent factors (and discarding the rest). In the context of term-document matrices, these methods often claim (implicitly or explicitly) to generalize over individual terms to identify the topics, semantics, or meaning.[3] Methods that fall generally under this category include: factor analysis, principal component analysis, independent components analysis, latent semantic analysis (LSA) or indexing, nonnegative matrix factorization, probabilistic latent semantic analysis, iterative scaling, and latent Dirichlet allocation. Some of these methods are very closely related; for the remainder of this subsection, I look at two extremely popular (and reasonably distinct) methods that will be important to the proceeding discussion: latent semantic analysis (LSA) and latent Dirichlet allocation (LDA).

LSA (Landauer and Dumais, 1997) relies on an important result from linear algebra, singular value decomposition (Golub and Van Loan, 1996), which allows that any positive matrix X

[3] In the context of distributional bag-of-word approaches, I do not think there is an important distinction to be drawn between *topic*, *semantics*, and *meaning*; I will continue to use prefer the first of these terms.

can be decomposed into the product of three matrices, U, Σ, and V^T. U and V^T are orthonormal matrices which provide new formulations of the rows and columns of the original matrix, while the values in the diagonal matrix Σ are the singular values, which indicate the importance of each column (dimension) of U with respect to the amount of variation in the original matrix that has been captured; from left to right each singular value is less than or equal to the last. The particulars of how the decomposition is achieved are beyond the scope of this discussion.[4] If we consider a version of Σ, Σ_k, where all but the first k singular values have been set to zero, the matrix $X_k = U\Sigma_k V^T$ is the best k rank approximation, with respect to the Frobenius norm, of the original matrix X. For LSA, depending on whether terms or documents are of interest, either U (for terms) or V (for documents) is used. For our purposes here, our LSA (term) vectors are the k dimensional truncations of $U\Sigma_k$, that is the U matrix weighed by the first k singular values; under this formulation, LSA is identical to dimensionality reduction via principal components analysis (PCA),[5] though we will prefer the term LSA since it is specific to term-document matrices. The final step of LSA is typically to compare LSA vectors by means of cosine similarity. Note that choosing the appropriate value of k for a given task in a given corpus is not trivial; rigorous empirical testing to find the optimal k is preferred (Deerwester et al., 1990).

LDA (Blei et al., 2003) is an example of a hierarchical Bayesian model, which, in general terms, views a collection of data (e.g. texts) as the direct result of a series of probabilistic choices, which in turn are conditioned by a set of parameters; both the probabilities and (when desired) the parameters can be estimated from the data using posterior inference, i.e. Bayesian reasoning. Under LDA, a text is viewed as a sample from a probability distribution over a set of topics; the distribution is selected from a Dirichlet distribution with prior α, and then each word

[4] Much of the focus in relevant work is how to do decomposition quickly while preserving its desirable properties. There are various software packages that include efficient SVD calculation.

[5] Though it is customary when applying PCA to center each column around the mean, a step which, in practice, is undesirable in extremely high-dimensional matrices such as those we use here, since it makes them much denser.

in the text is generated by first selecting a topic from the (text) probability distribution, and then a word from the word distribution for each topic ϕ; the topic-word distributions are also drawn from a Dirichlet, with prior β. Again, we do not pursue the technical details of estimating the desired probability distributions, though we note that there are two general approaches to the inference task: optimize a more tractable approximation, as in variational Bayes, or use a Markov chain Monte Carlo method such as (collapsed) Gibbs sampling (Porteous et al., 2008). Though the Dirichlet is the most popular distribution used for Bayesian topic modeling, other distributions such as the logistic normal distribution can be used to achieve other effects, for instance a correlation among different topics (Blei and Lafferty, 2007). Under the correlated topic model, the co-variance of the topics is controlled by another parameter to the model, the matrix Σ.

Finally, there has been some recent interest in exploring the extent to which the 'topics' identified by probabilistic topic modeling actually correspond to real human topics (Chang et al., 2009), and whether the judgments of humans can be modeled automatically (Newman et al., 2010). Chang (2009) presented Mechanical Turk workers with a pair of word choice tasks that rely on topic coherence, with the topics generated by one of three topic models: probabilistic LSA, LDA, and correlated topic models (CTMs). Interestingly, the traditional metric of how good a model is, i.e. higher likelihood on new data, actually correlated negatively with human judgments of coherence; CTMs had higher likelihood, but significantly lower human performance. They also found that a smaller number of topics generally resulted in more coherent topics. Newman (2010) tested a number of different metrics for automatically determining whether a topic (as derived by LDA) was coherent, as judged by humans; generally speaking, metrics based on WordNet were not useful, but they found that a measure based on pointwise mutual information (see discussion in next section) in Wikipedia correlated with human judgment nearly as well as other human judgments did. Follow-up work has focused

on regularizing LDA to produce more coherent topics using priors derived from external data (Newman et al., 2011).

3.3 Polarity lexicon induction

In this section, I review approaches to the automatic acquisition of polarity lexicons for the purposes of sentiment analysis. The reason I focus on sentiment lexicons (rather than synonym extraction, for instance) should be clear: The idea of a positive/negative spectrum is, on the surface, very similar to the notion of stylistic dimensions (or clines), and, as we will see, some of the methodology is directly applicable.

I will first briefly summarize methods based primarily on lexicographical resources such as WordNet. The standard approach for any method is to begin with a small set of seed terms. In WordNet, the polarity of new terms can be then be predicted from the association between seed terms, as measured by path distance (Kamps et al., 2004). The most well-known WordNet-based method is SentiWordNet (Esuli and Sebastiani, 2006), which uses semantic relationships such as synonymy and antonymy to expand the seed set, and then takes these as training instances to build classifiers to identify positive, negative, and neutral synsets based on their gloss; the latest version (Baccianella et al., 2010) includes an additional iterative process to improve resulting labels via a random walk through the graph; Hassan and Radev (2010) also use a random walk model in WordNet, achieving state-of-the-art results. Other methods based on lexical resources include that of Takamura et al. (2005), a model which spreads polarity in a gloss-derived graph using methods from physics to predict electron spin; Rao and Ravichandra (2009), who test the effectiveness of mincuts and label propagation in the WordNet graph; and Mohammad et al. (2009), who expand their seed sets using synonyms in a thesaurus and common affixes. Though some of the graph-based methods are applicable to the task of stylistic lexicon creation, as argued earlier it appears that stylistic variation is usually filtered out of

these kinds of resources (stylistic variants are placed under a single synset) and so I will not directly pursue acquisition using such resources.

The other major source of information for polarity dictionaries is corpora. Initial work by Hatzivassiloglou and McKeown (1997) used the Wall Street Journal corpus, and was only concerned with whether adjectives were positive or negative. The authors posited that the choice of connectives joining an adjective tends to indicate whether the two adjectives are of the same or opposing orientation. Using counts of adjective conjunctions, the authors derived a dissimilarity value for each pair of adjectives, and then used that to cluster the adjectives into two groups; neutral words were ignored. Another corpus-based method relies on specific linguistic patterns is that of Kaji and Kitsuregawa (2007); they, however, use a larger set of patterns and a much larger web corpus, deciding if a word is polar or not based on frequency of occurrence in polar sentences that contain positive or negative patterns.

Turney (2002) derives semantic orientation (SO) values for bigrams using pointwise mutual information (Church and Hanks, 1990) of the phrase and two seed words of opposing polarity ("excellent" and "poor"), calculated using hit counts (using the now-defunct AltaVista search engine and its NEAR operator). Briefly, PMI measures the extent to which the joint probability of a pair of events (or outcomes) varies from what we would expect based only on their individual marginal probabilities; high PMI means that two events happen together far more often than would be expected by chance. Turney and Littman (2003) use a slightly modified form of this same algorithm to calculate the SO of individual words, expanding their set of seed words to include seven of each polarity. They also test another measurement of relatedness based on LSA; the polarity of words is based on the cosine similarity of the LSA-derived word vectors of each word to the seed terms. In general, LSA outperformed PMI, even though LSA was necessarily based on much smaller corpora. A serious problem with using the hit counts to calculate SO is that the internet is constantly in flux. Taboada et al. (2006) report that the

Google search engine (with its text-wide AND operator) is much less reliable for the task of calculating SO-PMI; the SO values of adjectives calculated using hit counts from the Google API varied widely from day to day.

Velikovich et al. (2010) combine large corpora with a graph-based approach. They assign polarity scores to *n*-grams on the basis of the maximum weighed path from the phrase to seed terms. Rather than using a lexicon such as WordNet, the connection weights between phrases in the graph are derived using the similarity of context vectors (from a 6-word context) aggregated over all mentions of the *n*-gram in 4 billion web documents. They argue that, although many of these edges are unreliable, using maximized path weight rather than some form of label propagation limits the effect of these errors; their method outperforms another system based on label propagation in WordNet.

3.4 Formality lexicon induction

3.4.1 Introduction

The formality of a word relates to its appropriateness in a given context.[6] Consider, for example, the problem of choice among near-synonyms: there are only minor denotational differences among synonyms such as *get*, *acquire*, *obtain*, and *snag*, but it is difficult to construct a situation where any choice would be equally suitable. The key difference between these words is their formality, with *acquire* the most formal and *snag* the most informal.

We conceive of formality as a continuous property. This approach is inspired by writing assistance resources such as *Choose The Right Word* (Hayakawa, 1994), in which differences

[6]The work presented in this section is adapted from two publications: "Inducing lexicons of formality from corpora" by Julian Brooke, Tong Wang, and Graeme Hirst, published in *Proceedings of the 7th International Conference on Language Resources and Evaluation, Workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods* (Brooke et al., 2010b); and "Automatic acquisition of lexical formality" by Julian Brooke, Tong Wang, and Graeme Hirst, published in the *Proceedings of the 23rd International Conference on Computational Linguistics* (Brooke et al., 2010a).

between synonyms are generally described in relative rather than absolute terms, as well as linguistic literature in which the quantification of stylistic differences among genres is framed in terms of dimensions rather than discrete properties (Biber, 1988). We begin by defining the *formality score* (FS) for a word as a real number value in the range 1 to -1, with 1 representing an extremely formal word, and -1 an extremely informal word. A formality lexicon, then, gives an FS to every word within its coverage.

3.4.2 Data and resources

We begin with two word lists, one formal and one informal, that we use both as seeds for our lexicon construction methods and as test sets for evaluation (our gold standard). We assume that all slang terms are by their very nature informal and so our 138 informal seeds were taken primarily from an online slang dictionary[7] (e.g. *wuss*, *grubby*) and also include some contractions and interjections (e.g. *cuz*, *yikes*). The 105 formal seeds were selected from a list of discourse markers (e.g. *moreover*, *hence*) and adverbs from a sentiment lexicon (e.g. *preposterously*, *inscrutably*); these sources were chosen to avoid words with overt topic, and to ensure that there was some balance of sentiment across formal and informal seed sets. Part of speech, however, is not balanced across our seed sets.

Another test set we use to evaluate our methods is a collection of 399 pairs of near-synonyms from *Choose the Right Word* (CTRW); each pair was either explicitly or implicitly compared for formality in the book. Implicit comparison included statements such as *this is the most formal of these words*; in those cases, and more generally, we avoided words appearing in more than one comparison (there are no duplicate words in our CTRW set), as well as multiword expressions and words whose formality is strongly ambiguous (i.e. word-sense dependent). An example of this last phenomenon is the word *cool*, which is used colloquially in

[7] http://onlineslangdictionary.com/

the sense of *good* but more formally as in the sense of *cold*. Partly as a result of this polysemy, which is clearly more common among informal words, our pairs are biased toward the formal end of the spectrum; although there are some informal comparisons, e.g. *bellyache/whine*, *wisecrack/joke*, more typical pairs include *determine/ascertain* and *hefty/ponderous*. Despite this imbalance, one obvious advantage of using near-synonyms in our evaluation is that factors other than linguistic formality (e.g. topic, opinion) are less likely to influence performance. In general, the CTRW allows for a more objective, fine-grained evaluation of our methods, and is oriented towards our primary interest, near-synonym word choice.

To test the performance of our semi-supervised method beyond English, a native speaker of Mandarin Chinese created two sets of Chinese two-character words, one formal, one informal, based on but not limited to the words in the English sets. The Chinese seeds include 49 formal seeds and 43 informal seeds.

Our corpora fall generally into three categories: formal (written) copora, informal (spoken) corpora, and mixed corpora. The Brown Corpus (Francis and Kučera, 1982), our development corpus, is used here both as a formal and mixed corpus. Although extremely small by modern corpus standards (only 1 million words), the Brown Corpus has the advantage of being compiled explicitly to represent a range of American English, though it is all of the published, written variety. The Switchboard (SW) Corpus is a collection of American telephone conversations (Godfrey et al., 1992), which contains roughly 2400 conversations with over 2.6 million word tokens; we use it as an informal counterpart to the Brown Corpus. Like the Brown Corpus, The British National Corpus (Burnard, 2000) is a manually-constructed mixed-genre corpus; it is, however, much larger (roughly 100 million words). It contains a written portion (90%), which we use as a formal corpus, and a spontaneous spoken portion (4.3%), which we use as an informal corpus. Our other mixed corpora are two blog collections available to us: the first, which we call our development blog corpus (Dev-Blog) contains a total of over 900,000

English blogs, with 216 million tokens.[8] The second is the 'first tier' English blogs included in the publicly available ICSWM 2009 Spinn3r Dataset (Burton et al., 2009), a total of about 1.3 billion word tokens in 7.5 million documents. For our investigations in Chinese, we use the Chinese portion of the ICSWM blogs, approximately 25.4 million character tokens in 86,000 documents.

3.4.3 Basic methods

The simplest kind of formality measure is based on word length, which is often used directly as an indicator of formality for applications such as genre classification (Karlgren and Cutting, 1994). Here, we use logarithmic scaling to derive an FS based on word length. Given a maximum word length L[9] and a word w of length l, the formality score function, $FS(w)$, is given by:

$$FS(w) = -1 + 2\frac{\log l}{\log L}$$

For hyphenated terms, the length of each component is averaged. Though this metric works relatively well for English, we note that it is problematic in a language with significant word agglutination (e.g. German) or without an alphabet (e.g. Chinese, see below).

Another straightforward method is the assumption that Latinate prefixes and suffixes are indicators of formality in English (Kessler et al., 1997), i.e. informal words will not have Latinate affixes such as *-ation* and *intra-*. Here, we simply assign words that appear to have such a prefix or suffix an FS of 1, and all other words an FS of -1.

Our frequency methods derive FS from word counts in corpora. Our first, naive approach assumes a single corpus, where either formal words are common and informal words are rare,

[8] These blogs were gathered by the University of Toronto Blogscope project (www.blogscope.net) over a week in May 2008.

[9] We use an upper bound of 28 characters, which is the length of *antidisestablishmentarianism*, the prototypical longest word in English; this value of L provides an appropriate formality/informality threshold, between 5- and 6-letter words

or vice versa. To smooth out the Zipfian distribution, we use the frequency rank of words as exponentials; for a corpus with R frequency ranks, the FS for a word of rank r under the *formal is rare* assumption is given by:

$$FS(w) = -1 + 2\frac{e^{(r-1)}}{e^{(R-1)}}$$

Under the *informal is rare* assumption:

$$FS(w) = 1 - 2\frac{e^{(r-1)}}{e^{(R-1)}}$$

A more sophisticated method is to use two corpora that are known to vary with respect to formality and use the relative appearance of words in each corpus as the metric. If word appears n times in a (relatively) formal corpus and m times in an informal corpus (and one of m, n is not zero), we derive:

$$FS(w) = -1 + 2\frac{n}{m \times N + n}$$

Here, N is the ratio of the size (in tokens) of the informal corpus (IC) to the formal corpus (FC). We need the constant N so that an imbalance in the size of the corpora does not result in an equivalently skewed distribution of FS.

The rest of our simple methods rely on co-occurrence, based on some metric of association. One such metric is pointwise mutual information (PMI); we derive probabilities using a word versus document matrix, with the FS of each word calculated as follows:

$$FS(w) = \frac{1}{N}\left(\sum_{f \in F}\frac{P(w,f)}{P(w)P(f)} - \sum_{i \in I}\frac{P(w,i)}{P(w)P(i)}\right)$$

Here, F is the list of formal seeds, I is the list of informal seeds, and N is a normalization factor, either $argmax|FS'(w_F)|$ (for all w $FS'(w) > 0$) or $argmax|FS'(w_I)|$ (for all w, $FS'(w) < 0$), where $FS'(w)$ is the calculation before normalization; this last insures that the FS will

be the range 1 to -1. $P(w, f)$ is the probability (the count) of the word appearing with a particular formal seed in the same document. Note that calculating PMI typically involves taking the logarithm, but to avoid having to deal with $-\infty$ when the joint probability is zero (which happens often) we skip this step.

Our next method is LSA, already discussed in some detail in Section 3.2. Besides the choice of dimensionality, k, another factor is the size of a passage, which could be as large as a full document or as small as a sentence; here, we consider documents and paragraphs as possible passages.[10] A third variable that we investigated is the weighting of values in the original matrix; Turney and Littman, for instance, used *tf·idf*, however it was not clear that this was appropriate for our task, and so we tested various possible options (binary, *tf*, *idf*, and *td·idf*). We also consider the effect of lemmatization.

LSA is computationally intensive; in order to apply it to extremely large blog corpora, we need to filter the documents and terms before building our term–document matrix. We adopt the following strategy: to limit the number of documents in our term–document matrix, we first remove documents less than 100 tokens in length, with the rationale that these documents provide less co-occurrence information. Second, we remove documents that either do not contain any target words (i.e. one of our seeds or CTRW test words), or contain only target words which are among the most common 20 in the corpus; these documents are less likely to provide us with useful information, and the very common target terms will be well represented regardless. We further shrink the set of terms by removing all hapax legomena; a single appearance in a corpus is not enough to provide reliable co-occurrence information, and roughly half the types in our blog corpora appear only once. Finally, we remove symbols and all words which are not entirely lower case; we are not interested, for instance, in numbers, acronyms, and proper nouns. We can estimate the effect this filtering has on performance by testing it both ways in a

[10] Preliminary testing with sentences suggested that the resulting matrices were far too sparse to be useful; we omit those results here.

development corpus.

Once a k-dimensional vector for each relevant word is derived using LSA, a standard method is to use the cosine of the angle between a word vector and the vectors of seed words to identify how similar the distribution of the word is to the distribution of the seeds. To begin, each formal seed is assigned an FS value of 1, each informal seed an FS value of -1, and then a raw seed similarity score (FS') is calculated for each word w:

$$FS'(w) = \sum_{s \in S, s \neq w} W_s \times FS(s) \times \cos(\theta(\mathbf{w}, \mathbf{s}))$$

S is the set of all seeds. Note that seed terms are excluded from their own FS calculation, this is equivalent to *leave-one-out* cross-validation. W_s is a weight that depends on whether s is a formal or informal seed, W_i (for informal seeds) is calculated as:

$$W_i = \frac{\sum_{f \in F} FS(f)}{|\sum_{i \in I} FS(i)| + \sum_{f \in F} FS(f)}$$

and W_f (for formal seeds) is:

$$W_f = \frac{|\sum_{i \in I} FS(i)|}{|\sum_{i \in I} FS(i)| + \sum_{f \in F} FS(f)}$$

Here, I is the set of all informal seeds, and F is the set of all formal seeds. These weights have the effect of countering any imbalance in the seed set, as formal and informal seeds ultimately have the same (potential) influence on each word, regardless of their count. This weighting is necessary for the iterative extension of this method discussed in the next section.

We calculate the final FS as follows:

$$FS(w) = \frac{FS'(w) - FS'(r)}{N_w}$$

The word r is a reference term, a common function word that has no formality.[11] This has the effect of countering any (moderate) bias that might exist in the corpus; in the Brown Corpus, for instance, function words have positive formality before this step, simply because formal words occurred more often in the corpus. N_w is a normalization factor, either

$$N_w = \max_{w_i \in I'} |FS'(w_i) - FS'(r)|$$

for all $w_i \in I'$ or

$$N_w = \max_{w_f \in F'} |FS'(w_f) - FS'(r)|$$

for all $w_f \in F'$. I' contains all words w such that $FS'(w) - FS'(r) < 0$, and F' contains all words w such that $FS'(w) - FS'(r) > 0$. This ensures that the resulting lexicon has terms exactly in the range 1 to -1, with the reference word r at the midpoint.

Another method that is available to us, due to the relatively large size of our seed sets, is derivation of FS by means of regression, using machine learning algorithms. We speculate that this might be preferable to the cosine method since the irrelevant dimensions might be discarded from the model, whereas in the cosine calculations these dimensions would show up as noise. To investigate the effectiveness of this approach, we tested various regression algorithms included in the WEKA software suite (Witten and Frank, 2005); below, we present results for two, linear regression and Gaussian processes, which performed well based on the r-squared value with 10-fold cross-validation; for both we used the default settings for WEKA (version 3.6.2), which for Gaussian processes entails a classifier with an RBF kernel. Training was carried out using the k-dimensional vectors of our formal and informal seeds; for the purposes of training the former were assigned a value of 1, the latter -1. Since the model

[11] The particular choice of this word is relatively unimportant; common function words all have essentially the same LSA vectors because they appear at least once in nearly every document of any size. For English, we chose r = *and*, and for Chinese, r = *yinwei* (*because*); there does not seem to be an obvious two-character, formality-neutral equivalent to *and* in Chinese.

applied to new data could potentially fall outside that range, appropriate normalization of the output is also necessary in this case.

We also tested the LSA method in Chinese. The only major relevant difference between Chinese and English is word segmentation: Chinese does not have spaces between words. To sidestep this problem, we simply included all character bigrams found in our corpus. The drawback of this approach in the inclusion of a huge number of nonsense 'words' (1.3 million terms in just 86,000 documents), however we are at least certain to identify all instances of our seeds.

3.4.4 Hybrid methods

There are a number of ways to leverage the information we derive from our basic methods. One intriguing option is to use the basic FS measures as the starting point for an iterative process using the LSA cosine similarity. Under this paradigm, all words in the starting FS lexicon are potential seed words; we choose a cutoff value for inclusion in the seed word set (e.g. words which have at least .5 or $-.5$ FS), and then carry out the cosine calculations, as above, to derive new FS values (a new FS lexicon). We can repeat this process as many times as required, with the idea that the connections between various words (as reflected in their LSA-derived vectors) will cause the system to converge towards the true FS values.

A simple hybrid method that combines the two word count models uses the ratio of word counts in two corpora to define the center of the FS spectrum, but single corpus methods to define the extremes. Formally, if m and n (word counts for the informal corpus IC and formal corpus FC, respectively) are both non-zero, then FS is given by:

$$FS(w) = -0.5 + \frac{n}{m \times N + n}$$

However, if n is zero, FS is given by:

$$FS(w) = -1 + 0.5\frac{e^{\sqrt{r_{IC}-1}}}{e^{\sqrt{R_{IC}-1}}}$$

where r_{IC} is the frequency rank of the word in IC, and R_{IC} is the total number of ranks in IC. If m is zero, FS is given by:

$$FS(w) = 1 - 0.5\frac{e^{\sqrt{r_{FC}-1}}}{e^{\sqrt{R_{FC}-1}}}$$

where i is the rank of the word in IC, and R_{IC} is the total number of frequency ranks in IC). This function is undefined in the case where m and n are both zero. Intuitively, this is a kind of backoff, relying on the idea that words of extreme formality are rare even in a corpus of corresponding formality, whereas words in the *core vocabulary* (Carter, 1998), which are only moderately formal, will appear in all kinds of corpora, and thus are amenable to the ratio method.

Finally, we explore a number of ways to combine lexicons directly. The motivation for this is that the lexicons have different strengths and weaknesses, representing partially independent information. An obvious method is an averaging or other linear combination of the scores, but we also investigate vote-based methods (requiring agreement among n dictionaries). Beyond these simple options, we test support vector machines and naive Bayes classification using the WEKA software suite (Witten and Frank, 2005), applying 10-fold cross-validation using default WEKA settings for each classifier. The features here are task dependent; for the pairwise task, we use the difference between the FS value of the words in each lexicon, rather than their individual scores. Finally, we can use the weights from the SVM model of the CTRW (pairwise) task to interpolate an optimal formality lexicon.

3.4.5 Evaluation

We evaluate our methods using the gold standard judgments from the seed sets and CTRW word pairs. To differentiate the two, we continue to use the term *seed* for the former; in this context, however, these 'seed sets' are being viewed as a test set (recall that our LSA method is equivalent to *leave-one-out* cross-validation).

We derive the following measures: first, the coverage (Cov.) is the percentage of words in the set that are covered under the method. The class-based accuracy (C-Acc.) of our seed sets is the percentage of covered words which are correctly classified as formal (FS > 0) or informal (FS < 0). The pair-based accuracy (P-Acc.) is the result of exhaustively pairing words in the two seed sets and testing their relative formality; that is, for all $w_i \in I$ and $w_f \in F$, the percentage of w_i/w_f pairs where $FS(w_i) < FS(w_f)$. For the CTRW pairs there are only two metrics, the coverage and the pair-based accuracy; since the CTRW pairs represent relative formality of varying degrees, it is not possible to calculate a class-based accuracy.

Since there are a large range of options to consider, we decompose our evaluation into two steps.[12] In the first step (Table 3.2), we restrict ourselves to the smaller Brown corpus, but test a wider range of options, particularly related to LSA. In the second step (Table 3.3), we use a fixed set of the best options for LSA and focus on the benefits of larger corpora and hybridization with other feature information.

The results for the first (Brown-focused) evaluation are shown in Table 3.2; the numbers in parentheses below indicate the corresponding line of the table. In the first section of the table, the baseline provided by the word length (1) is quite high, particularly for seed set pairwise accuracy, indicating that nearly all the informal seed words are shorter than the formal seed words. Word length is not as effective with the fine-grained differences, however, and the class-based accuracy is low, as many formal seeds are incorrectly labeled as informal using

[12] These steps correspond roughly to the separate publications which originally presented this work.

Table 3.2: Seed coverage (%), class-based accuracy (%), pairwise accuracy (%), CTRW coverage (%) and pairwise accuracy (%) for various FS lexicon creation methods. Co-occurence defaults are Brown, no lemmatization, binary features, and document-level context.

	Seed set			CTRW set	
Method	**Cov.**	**C-Acc.**	**P-Acc.**	**Cov.**	**P-Acc.**
Baseline methods					
(1) Word length	100	74.9	91.8	100	63.7
(2) Latinate affixes	100	74.5	46.3	100	32.6
Word count methods					
(3) Word Counts, Brown, informal is rare,	51	63.7	68.3	59.6	18.5
(4) Word Counts, Brown, formal is rare	51	36.3	19.5	59.6	55.0
(5) Ratio, Brown and Switchboard	38	81.5	85.7	35.6	78.2
Co-occurrence methods					
(6) PMI, Brown	51.0	80.6	84.4	59.6	73.2
(7) LSA (k=100), cosine	51.0	88.7	96.1	59.6	53.8
(8) LSA (k=10), cosine	51.0	88.7	95.0	59.6	66.4
(9) LSA (k=3), cosine	51.0	89.5	94.5	59.6	73.9
(10) LSA (k=3), cosine, lemma	51.0	88.5	94.4	59.6	70.5
(11) LSA (k=100), cosine, paragraph	51.0	83.1	96.6	59.6	53.8
(12) LSA (k=10), cosine, paragraph	51.0	83.1	95.0	59.6	61.8
(13) LSA (k=3), cosine, paragraph	51.0	83.1	91.7	59.6	73.5
(14) LSA (k=3), *tf*, cosine	51.0	66.1	74.9	59.6	49.2
(15) LSA (k=3), *idf*, cosine	51.0	55.6	57.7	59.6	52.5
(16) LSA (k=3), *td*·*idf*, cosine	51.0	54.8	39.7	59.6	52.5
(17) LSA (k=100), Gaussian	51.0	71.8	83.8	59.6	38.2
(18) LSA (k=10), Gaussian	51.0	81.5	92.3	59.6	56.3
(19) LSA (k=3), Gaussian	51.0	87.1	92,7	59.6	56.7
(20) LSA (k=100), linear	51.0	58.9	57.6	59.6	53.4
(21) LSA (k=10), linear	51.0	79.0	88.9	59.6	58.4
(22) LSA (k=3), linear	51.0	75.8	86.8	59.6	61.8

our linear method. It is clear from the class-based accuracy score that Latinate suffixes and prefixes (2) are indicative of formality; they do not, however, provide information that allows for relative, more fine-grained distinctions. The advantage of these methods, of course, is their coverage.

The first two results in the second part of Table 3.2 (3–4) show that neither assumption (i.e. that formal words are rare or that informal words are rare) is particularly successful, though

they fail in different ways that are indicative of the formality make-up of the corpus and the test sets. Since the Brown corpus is a corpus of published written texts, and therefore more formal, the *informal is rare* hypothesis (3) is a better one for the extreme seed sets; however, in the CTRW test sets, which is more indicative of the formal end of the spectrum, this assumption fails spectacularly, with the model performing much worse than chance. The opposite is true for the *formal is rare* model (4), since it makes opposite predictions. Neither is directly useful for the task as a whole. Much better is the word ratio model using the Brown corpus as the formal dictionary and the Switchboard corpus as the informal dictionary (5); although the coverage is quite low, the score for pairwise accuracy in the CTRW set is the highest in Table 3.2, and the scores for the seed test are also quite good.

The co-occurrence results are presented in the third part of Table 3.2. The PMI results (6) are quite promising, given the simple nature of the calculation, though the LSA results (7–9) are better, particularly when the optimal value of k is used (9). To find that value, we tested all values between 1 and 10, and at intervals of 10 thereafter.

Looking at the options for LSA, lemmatization (10) has a small but consistently negative effect. More notable is the drop in performance when paragraphs rather than documents are taken as the unit in our word–passage matrix (11–13), suggesting that a *one level of formality per document* assumption is a relatively good one; the pairwise accuracy in the seed sets, though, is consistently high. With respect to weights, our original intuition was that a binary feature for appearance in a document was the best way to approach the construction of a word-document matrix; intuitively, there does not seem to be useful information that can be gleaned from the number of appearances of a formal or informal word in a document, nor should a word be weighted solely based on its rarity in a corpus. Indeed, our results (14–16) confirm this; applying *td·idf* or either of its component results in a major drop in performance across the board.

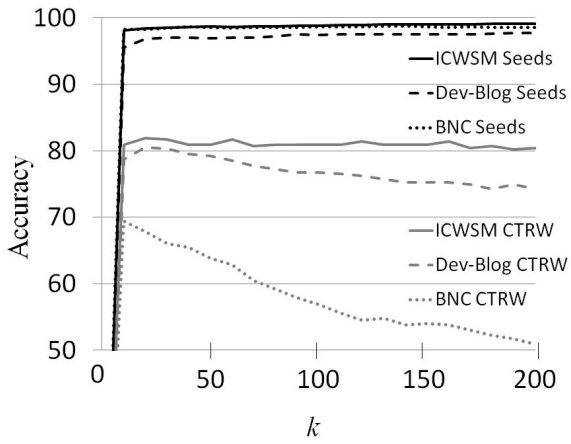


Figure 3.1: Seed and CTRW pairwise accuracy, LSA method for large corpora k, $10 \leq k \leq 200$.

Finally, we look at the results using machine learning regression methods rather than cosine distance to derive FS (17–22). Neither of the algorithms performs well on the CTRW set, with the Gaussian Processes method (22–24) particularly poor, despite its relative sophistication; one explanation is that it tries to maximize the extreme cases, failing on the more-subtle word distinctions. The performance differences related to increases in k are consistent with cosine but more marked, revealing themselves in all three accuracy measures, though with a great deal more variation across the methods.

Table 3.3 contains the results for the larger corpus and hybrid methods. The first two sections of Table 3.3 include results for simple methods: these are mostly the same as in Table 3.2 though (4) includes the word count ratio of written to spoken words in the BNC, which provides better coverage though not better performance in the CTRW set. The LSA results in Table 3.3 are the best for each corpus across the k values we tested. When both coverage and accuracy are considered, there is a clear benefit associated with increasing the amount of data, though the difference between the Dev-Blog and ICWSM suggests diminishing returns. The performance of the filtered Dev-Blog is actually slightly better than the unfiltered versions (though there is a drop in coverage), suggesting that filtering is a good strategy.

Table 3.3: Seed coverage, class-based accuracy, pairwise accuracy, CTRW coverage, and pairwise accuracy for various FS lexicons and hybrid methods (%).

Method	Seed set			CTRW set	
	Cov.	**C-Acc.**	**P-Acc.**	**Cov.**	**P-Acc.**
Simple					
(1) Word length	100	86.4	91.8	100	63.7
(2) Latinate affix	100	74.5	46.3	100	32.6
(3) Count ratio, Brown and Switchboard	38.0	81.5	85.7	36.0	78.2
(4) Count ratio, BNC Written vs. Spoken	60.9	89.2	97.3	38.8	74.3
(5) LSA (k=3), Brown	51.0	87.1	94.2	59.6	73.9
(6) LSA (k=10), BNC	94.7	83.0	98.3	96.5	69.4
(7) LSA (k=20), Dev-Blog	100	91.4	96.8	99.0	80.5
(8) LSA (k=20), Dev-Blog, filtered	99.0	92.1	97.0	97.7	80.5
(9) LSA (k=20), ICWSM, filtered	100	93.0	98.4	99.7	81.9
Hybrid					
(10) BNC ratio with backoff (4)	97.1	78.8	75.7	97.0	78.8
(11) Combined ratio with backoff (3 + 4)	97.1	79.2	79.9	97.5	79.9
(12) BNC weighted average (10,6), ratio 2:1	97.1	83.5	90.0	97.0	83.2
(13) Blog weighted average (9,7), ratio 4:1	100	93.8	98.5	99.7	83.4
(14) Voting, 3 agree (1, 6, 7, 9, 11)	92.6	99.1	99.9	87.0	91.6
(15) Voting, 2 agree (1, 11, 13)	86.8	99.1	100	81.5	96.9
(16) Voting, 2 agree (1, 12, 13)	87.7	98.6	100	82.7	97.3
(17) SVM classifier (1, 2, 6, 7, 9, 11)	100	97.9	99.9	100	84.2
(18) Naive Bayes classifier (1, 2, 6, 7, 9, 11)	100	97.5	99.8	100	83.9
(19) SVM (Seed) weights (1, 2, 6, 7, 9, 11)	100	98.4	99.8	100	80.5
(20) SVM (CTRW) weights (1, 6, 7, 9, 11)	100	93.0	99.0	100	86.0
(21) Average (1, 6, 7, 9, 11)	100	95.9	99.5	100	84.5

When testing in the Brown, we noted that CTRW set performance in the Brown dropped for $k > 3$, while performance on the seed set was mostly steady as k increased. Figure 3.1 shows the pairwise performance of each test set for the larger corpora across various k. The results here are similar; all three corpora reach a CTRW maximum at a relatively low k values (though higher than Brown Corpus); however the seed set performance in each corpus continues to improve (though marginally) as k increases, while CTRW performance drops. An explanation for this is that the seed terms represent extreme examples of formality; thus there are numerous semantic dimensions to distinguish them. However, the CTRW set includes near-synonyms,

many with only relatively subtle differences in formality; for these pairs, it is important to focus on the core dimensions relevant to formality, which are among the first discovered in a factor analysis of mixed-register texts (Biber, 1988).

With regards to hybrid methods, we first briefly summarize our testing with the iterative model, which included extensive experiments using basic lexicons and the LSA vectors derived from the Brown Corpus, and some targeted testing with the blog corpora (iteration on these corpora is extraordinarily time-consuming). In general, we found only that there were only small, inconsistent benefits to be gained from the iterative approach. More generally, the intuition behind the iterative method, i.e. that performance would increase with a drastic increase in the number of seeds, was found to be flawed: in other testing, we found that we could randomly remove most of the seeds without negatively affecting performance. Even at relatively high k values, it seems that a few seeds are enough to calibrate the model.

The ratio (with backoff) hybrid built from the BNC (10) provides CTRW performance that is comparable to the best LSA models, though performance in the seed sets is somewhat poor; supplementing with word counts from the Brown Corpus and Switchboard Corpus provides a small improvement (11). The weighed hybrid dictionaries in (12,13) demonstrate that it is possible to effectively combine lexicons built using two different methods on the same corpus (12) or the same method on different corpora (13); the former, in particular, provides an impressive boost to CTRW accuracy, indicating that word count and word association methods are partially independent.

The remainder of Table 3.3 shows the best results using voting, averaging, and weighting. The voting results (14–16) indicate that it is possible to sacrifice some coverage for very high accuracy in both sets, including a near-perfect score in the seed sets and significant gains in CTRW performance. In general, the best accuracy without a significant loss of coverage came from 2 of 3 voting (15–16), using dictionaries that represented our three basic sources of in-

formation (word length, word count, and word association). The machine learning hybrids (17–18) also demonstrate a marked improvement over any single lexicon, though it is important to note that each accuracy score here reflects a different task-specific model. Hybrid FS lexicons built with the weights learned by the SVM models (19–20) provide superior performance on the task corresponding to the model used, though the simple averaging of the best dictionaries (21) also provides good performance across all evaluation metrics.

Finally, the LSA results for Chinese are modest but promising, given the relatively small scale of our experiments: we saw a pairwise accuracy of 82.2%, with 79.3% class-based accuracy ($k = 10$). We believe that the main reason for the generally lower performance in Chinese (as compared to English) is the modest size of the corpus, though our simplistic character bigram term extraction technique may also play a role. As mentioned, smaller seed sets do not seem to be an issue. Interestingly, the class-based accuracy is 10.8% lower if no reference word is used to calibrate the divide between formal and informal, suggesting a rather biased corpus (towards informality); in English, by comparison, the reference-word normalization had a slightly negative effect on the LSA results, though the effect mostly disappeared after hybridization. The obvious next step is to integrate a Chinese word segmenter, and use a larger corpus. We could also try word count methods, though finding appropriate (balanced) resouces similar to the BNC might be a challenge; (mixed) blog corpora, on the other hand, are easily collected.

3.5 Readability lexicon induction

3.5.1 Introduction

With its goal of identifying documents appropriate to readers of various proficiencies, automatic analysis of readability is typically approached as a text-level classification task; we will

discuss this aspect later in Section 4.1.[13] Nevertheless, information about the relative difficulty of individual lexical items, in addition to being useful for text readability classification (Kidwell et al., 2009), can be applied to other tasks, for instance lexical simplification (Carroll et al., 1999; Burstein et al., 2007). This work was motivated by an interest in providing students with educational software that is sensitive to the difficulty of particular English expressions, providing proactive support for those which are likely to be outside a reader's vocabulary. However, our existing lexical resource is coarse-grained and lacks coverage, so we would like to expand it using automated methods. As with our discussion of formality in the previous section, a readability lexicon assigns a number to each word, here reflecting the relative degree of experience that might be expected before the average reader would be able to recognize the word (correctly) on sight.

3.5.2 Related work

Simple metrics form the basis of much (text-level) readability work: most involve linear combinations of word length, syllable count, and sentence length (Kincaid et al., 1975; Gunning, 1952), though the popular Dale-Chall reading score (Dale and Chall, 1995) is based on a list of 3000 'easy' words; a recent review suggests these metrics are fairly interchangeable (van Oosten et al., 2010). In terms of computational approaches, the work of Kidwell et al. (2009) is perhaps closest to our work here. Like the above, their goal is text readability classification, but they proceed by first deriving an age of acquisition for each word based on its statistical distribution in age-annotated texts. Also similar is the work of Li and Feng (2011), who are critical of raw frequency as an indicator and instead identify core vocabulary based on the common use of words across different age groups.

[13] The work presented in this section is based on "Building readability lexicons with unannotated corpora" by Julian Brooke, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst, published in the *Proceedings of the NAACL '12 Workshop on Predicting and Improving Text Readability* (Brooke et al., 2012b).

Table 3.4: Examples from the Difficulty lexicon

Beginner
coat, away, arrow, lizard, afternoon, rainy, carpet, earn, hear, chill
Intermediate
bale, campground, motto, intestine, survey, regularly, research, conflict
Advanced
contingency, scoff, characteristic, potent, myriad, detracted, illegitimate, overture

3.5.3 Resources

Our primary resource is an existing lexicon.[14]. This resource, which we will refer to as the Difficulty lexicon, consists of 15,308 words and expressions classified into three difficulty categories: beginner, intermediate, and advanced. Beginner, which was intended to capture the vocabulary of early elementary school, is an amalgamation of various smaller sources, including the Dolch list (Dolch, 1948). The intermediate words, which include words learned in late elementary and middle school, were extracted from Internet-published texts written by students at these grade levels, and then filtered manually. The advanced words began as a list of common words that were in neither of the original two lists, but they have also been manually filtered; they are intended to reflect the vocabulary understood by the average high school student. Table 3.4 contains some examples from each list.

For our purposes here, we only use a subset of the Difficulty lexicon: we filtered out inflected forms, proper nouns, and words with non-alphabetic components (including multiword expressions) and then randomly selected 500 words from each level for our test set and 300 different words for our development/training set. Rather than trying to duplicate our arbitrary three-way distinction by manual or crowdsourced means, we instead focused on the relative difficulty of individual words: for each word in each of the two sets, we randomly selected

[14]This lexicon was built under the supervision of one of the authors of the original paper, Frasier Shein.

three comparison words, one from each of the difficulty levels, forming a set of 4500 test pairs (2700 for the development set): $1/3$ of these pairs are words from the same difficulty level, $4/9$ are from adjacent difficulty levels, and the remaining $2/9$ are at opposite ends of our difficulty spectrum.

Our crowdsourced annotation was obtained using Crowdflower, which is an interface built on top of Mechanical Turk. For each word pair to be compared, we elicited 5 judgments from workers. Rather than frame the question in terms of difficulty or readability, which we felt was too subjective, we instead asked which of the two words the worker thought he or she learned first: the worker could choose either word, or answer "about the same time". They were instructed to choose the word they did know if one of the two words was unknown, and "same" if both were unknown. For our evaluation, we took the majority judgment as the gold standard; when there was no majority judgment, then the words were considered "the same". To increase the likelihood that our workers were native speakers of English, we required that the responses come from the US or Canada. Before running our main set, we ran several smaller test runs and manually inspected them for quality; although there were outliers, the majority of the judgments seemed reasonable.

Our corpus is the ICWSM Spinn3r 2009 dataset (Burton et al., 2009), which we have already seen was effective for formality lexicon creation. As in the earlier work, we use only the documents which have at least 100 tokens. The corpus has been tagged using the TreeTagger (Schmid, 1995).

3.5.4 Methods

Our method for lexicon creation involves first extracting a set of relevant numerical features for each word type. We can consider each feature as defining a lexicon on its own, which can be evaluated using our test set. Our features can be roughly broken into three types: simple

features, document readability features, and co-occurrence features. The first of these types does not require much explanation: it includes the length of the word, measured in terms of letters and syllables (the latter is derived using a simple but reasonably accurate vowel-consonant heuristic), and the log frequency count in our corpus.[15]

The second feature type involves calculating simple readability metrics for each document in our corpus, and then defining the relevant feature for the word type as the average value of the metric for all the documents that the word appears in. For example, if D_w is the set of documents where word type w appears and d_i is the ith word in a document d, then the *document word length* (DWL) for w can be defined as follows:

$$DWL(w) = |D_w|^{-1} \sum_{d \in D_w} \frac{\sum_{i=0}^{|d|} length(d_i)}{|d|}$$

Other features calculated in this way include: the document sentence length, that is the average token length of sentences; the document type-token ratio[16]; and the document lexical density, the ratio of content words (nouns, verbs, adjectives, and adverbs) to all words.

The co-occurrence features are a generalized form of the LSA method, already described for formality lexicon creation in the previous subsection. As before, after creating LSA vectors we select two sets of seed words (P and N) which will represent the ends of the spectrum which we are interested in deriving. We derive a feature value V for each word by summing the cosine similarity of the word vector with all the seeds:

$$V(\mathbf{w}) = \frac{\sum_{\mathbf{p} \in P} \cos(\theta(\mathbf{w}, \mathbf{p}))}{|P|} - \frac{\sum_{\mathbf{n} \in N} \cos(\theta(\mathbf{w}, \mathbf{n}))}{|N|}$$

We further normalize this to a range of 1 to -1, centered around the core vocabulary word

[15] Though it is irrelevant when evaluating the feature alone, the log frequency was noticeably better when combining frequency with other features.

[16] We calculate this using only the first 100 words of the document, to avoid the well-documented influence of length on TTR.

and. Here, we try three possible versions of P and N: the first, Formality, is the set of words used in our formality lexicon induction. The second, Childish, is a set of 10 common 'childish' concrete words (e.g. *mommy*, *puppy*) as N, and a set of 10 common abstract words (e.g. *concept*, *philosophy*) as P. The third, Difficulty, consists of the 300 beginner words from our development set as N, and the 300 advanced words from our development set as P. We tested several values of k for each of the seed sets (from 20 to 500); there was only small variation so here we just present our best results for each set as determined by testing in the development set.

Our final lexicon is created by taking a linear combination of the various features. We can find an appropriate weighting of each term by taking them from a model built using our development set. We test two versions of this: by default, we use a linear regression model where for training beginner words are tagged as 0, advanced words as 1, and intermediate words as 0.5. The second model is a binary SVM classifier; the features of the model are the difference between the respective features for each of the two words, and the classifier predicts whether the first or second word is more difficult. Both models were built using WEKA (Witten and Frank, 2005), with default settings except for feature normalization, which must be disabled in the SVM to get useful weights for the linear combination which creates our lexicon. In practice, we would further normalize our lexicon; here, however, this normalization is not relevant since our evaluation is based entirely on relative judgments. We also tested a range of other machine learning algorithms available in WEKA (e.g. decision trees and MaxEnt) but the crossvalidated accuracy was similar to or slightly lower than using a linear classifier.

3.5.5 Evaluation

All results are based on comparing the relative difficulty judgments made for the word pairs in our test set (or, more often, some subset) by the various sources. Since even the existing

Difficulty lexicon is not entirely reliable, we report agreement rather than accuracy. Except for agreement of Crowdflower workers, agreement is the percentage of pairs where the sources agreed as compared to the total number of pairs. For agreement between Crowdflower workers, we follow Taboada et al. (2011) in calculating agreement across all possible pairings of each worker for each pair. Although we considered using a more complex metric such as Kappa, we believe that simple pairwise agreement is in fact equally interpretable when the main interest is relative agreement of various methods; besides, Kappa is intended for use with individual annotators with particular biases, an assumption which does not hold here.

To evaluate the reliability of our human-annotated resources, we look first at the agreement within the Crowdflower data, and between the Crowdflower and our Difficulty lexicon, with particular attention to within-class judgments. We then compare the predictions of various automatically extracted features and feature combinations with these human judgments; since most of these involve a continuous scale, we focus only on words which were judged to be different.[17] For the Difficulty lexicon (Diff.), the n in this comparison is 3000, while for the Crowdflower (CF) judgments it is 4002.

3.5.6 Results

We expect a certain amount of noise using crowdsourced data, and indeed agreement among Crowdflower workers was not extremely high, only 56.6% for a three-way choice; note, however, that in these circumstances a single worker disagreeing with the rest will drop pairwise agreement in that judgement to 60%.[18] Tellingly, average agreement was relatively high (72.5%) for words on the extremes of our difficulty spectrum, and low for words in the same

[17] A continuous scale will nearly always predict some difference between two words. An obvious approach would be to set a threshold within which two words will be judged the same, but the specific values depend greatly on the scale and for simplicity we do not address this problem here.

[18] In 87.3% of cases, at least 3 workers agreed; in 56.2% of cases, 4 workers agreed, and in 23.1% of cases all 5 workers agreed.

difficulty category (46.0%), which is what we would expect. As noted by Taboada et al. (2011), when faced with a pairwise comparison task, workers tend to avoid the "same" option; instead, the proximity of the words on the underlying spectrum is reflected in disagreement. When we compare the crowdsourced judgements directly to the Difficulty lexicon, base agreement is 63.1%. This is much higher than chance, but lower than we would like, considering these are two human-annotated sources. However, it is clear that much of this disagreement is due to "same" judgments, which are three times more common in the Difficulty lexicon-based judgments than in the Crowdflower judgments (even when disagreement is interpreted as a "same" judgment). Pairwise agreement of non-"same" judgments for word pairs which are in the same category in the Difficultly lexicon is high enough (45.9%)[19] for us to conclude that this is not random variation, strongly suggesting that there are important distinctions within our difficulty categories, i.e. that it is not sufficiently fine-grained. If we disregard all words that are judged as same in one (or both) of the two sources, the agreement of the resulting word pairs is 91.0%, which is reasonably high.

Table 3.5 contains the agreement when feature values or a linear combination of feature values are used to predict the readability of the unequal pairs from the two manual sources. First, we notice that the Crowdflower set is obviously more difficult, probably because it contains more pairs with fairly subtle (though noticeable) distinctions. Other clear differences between the annotations: whereas for Crowdflower frequency is the key indicator, this is not true for our original annotation, which prefers the more complex features we have introduced here. A few features did poorly in general: syllable count appears too coarse-grained to be useful on its own, lexical density is only just better than chance, and type-token ratio performs at or below chance. Otherwise, many of the features within our major types give roughly the same performance individually.

[19] Random agreement here is 33.3%.

Table 3.5: Agreement (%) of automated methods with manual resources on pairwise comparison task (Diff. = Difficulty lexicon, CF = Crowdflower)

Features	Resource	
	Diff.	CF
Simple		
Syllable length	62.5	54.9
Word length	68.8	62.4
Term frequency	69.2	70.7
Document		
Avg. word length	74.5	66.8
Avg. sentence length	73.5	65.9
Avg. type-token ratio	47.0	50.0
Avg. lexical density	56.1	54.7
Co-occurrence		
Formality	74.7	66.5
Childish	74.2	65.5
Difficulty	75.7	66.1
Linear Combinations		
Simple	79.3	75.0
Document	80.1	70.8
Co-occurrence	76.0	67.0
Document+Co-occurrence	80.4	70.2
Simple+Document	87.5	79.1
Simple+Co-occurrence	86.7	78.2
All	**87.6**	**79.5**
All (SVM)	87.1	79.2

When we combine features, we find that simple and document features combine to positive effect, but the co-occurrence features are redundant with each other and, for the most part, the document features. A major boost comes, however, from combining either document or co-occurrence features with the simple features; this is especially true for our Difficulty lexicon annotation, where the gain is 7 to 8 percentage points. It does not seem to matter very much whether the weights of each feature are determined by pairwise classifier or by linear regression: this is interesting because it means we can train a model to create a readability spectrum with only pairwise judgments. Finally, we took all the 2500 instances where our two annotations agreed that one word was more difficult, and tested our best model against only

those pairs. Results using this selective test set were, unsurprisingly, higher than those of either of the annotations alone: 91.2%, which is roughly the same as the original agreement between the two manual annotations.

3.5.7 Discussion

Word difficulty is a vague concept, and we have admittedly sidestepped a proper definition here: instead, we hope to establish a measure of reliability in judgments of 'lexical readability' by looking for agreement across diverse sources of information. Our comparison of our existing resources with crowdsourced judgments suggests that some consistency is possible, but that granularity is, as we predicted, a serious concern, one which ultimately undermines our validation to some degree. An automatically derived lexicon, which can be fully continuous or as coarse-grained as needed, seems like an ideal solution, though the much lower performance of the automatic lexicon in predicting the more fine-grained Crowdflower judgments indicates that automatically-derived features are limited in their ability to deal with subtle differences. However, a visual inspection of the spectrum created by the automatic methods suggests that, with a judicious choice of granularity, it should be sufficient for our needs. In future work, we also intend to evaluate its use for readability classification, and perhaps expand it to include multiword expressions and syntactic patterns.

Our results clearly show the benefit of combining multiple sources of information to build a model of word difficulty. Word frequency and word length are of course relevant, and the utility of the document context features is not surprising, since they are merely a novel extension of existing proxies for readability. The co-occurrence features were also useful, though they seem fairly redundant and slightly inferior to document features; we posit that these features, in addition to capturing notions of register such as formality, may also offer semantic distinctions relevant to the acquisition process. For instance, children may have a large vo-

cabulary in very concrete domains such as animals, including words (e.g. *lizard*) that are not particularly frequent in adult corpora, while very common words in other domains (such as the legal domain) are completely outside the range of their experience. If we look at some of the examples which term frequency alone does not predict, they seem to be very much of this sort: *dollhouse*/*emergence*, *skirt*/*industry*, *magic*/*system*. Unsupervised techniques for identifying semantic variation, such as LSA, can capture these sorts of distinctions. However, our results indicate that simply looking at the readability of the texts that these sort of words appear in (i.e. our document features) is mostly sufficient, and less than 10% of the pairs which are correctly ordered by these two feature sets are different. In any case, an age-graded corpus is definitely not required.

There are a few other benefits of using word co-occurrence that we would like to touch on, though we leave a full exploration for future work. First, if we consider readability in other languages, each language may have different properties which render proxies such as word length much less useful (e.g. ideographic languages like Chinese or agglutinative languages like Turkish). However, word (or lemma) co-occurrence, like frequency, is essentially a universal feature across languages, and thus can be directly extended to any language, as we did for formality. Second, if we consider how we would extend difficulty-lexicon creation to the context of adult second-language learners, it might be enough to adjust our seed terms to reflect the differences in the language exposure of this population, i.e. we would expect difficulty in acquiring colloquialisms that are typically learned in childhood but are not part of the core vocabulary of the adult language.

3.6 Multi-dimensional Bayesian lexicon induction

3.6.1 Introduction

The formality lexicon discussed in Section 3.4 clearly captures some of the human intuitions about this stylistic dimension.[20] But in other cases, it fails. The word *cute*, for instance, is judged as extremely informal. Though it clearly belongs on the informal end of the spectrum, it is not slang, yet it is judged more informal than many slang terms. Why would that be? One possible explanation is that its extremity on the formality spectrum is actually due to the combined force of multiple stylistic dimensions; *cute*, I note, is a very subjective term and common in spoken language; it is not extremely informal, but it is extreme in other ways, ways that intuitively may be correlated with informality. LSA, with its independent, orthogonal dimensions, clearly reflects broad trends, but its tendency is to collapse, not distinguish, correlated variables. On the other hand, Bayesian topic models may offer the flexibility needed to separate out correlated stylistic dimensions, and thus build lexicons where the direct influence of correlated dimensions is minimized.

For this study we have chosen 3 dimensions (6 styles) which are clearly represented in the lexicon, which are discussed often in the relevant literature, and which fit well into the Leckie-Tarry conception of related subclines: colloquial vs. literary, concrete vs. abstract, and subjective vs. objective. In addition to a negative correlation between opposing styles, we also expect a positive correlation between stylistic aspects that tend toward the same main pole, situational (i.e. colloquial, concrete, subjective) or cultural (i.e. literary, abstract, objective). These correlations can potentially interfere with accurate lexical acquisition.

Besides our own work on formality (see Section 3.4) and the task of sentiment lexicon creation (see Section 3.3), we should note that there is other relevant lexical acquisition work,

[20]The work presented in the section is based on "A multi-dimensional Bayesian approach to lexical style" by Julian Brooke and Graeme Hirst, published in the *Proceedings of the 13th Annual Conference of the North American Chapter of the Association for Computational Lingusitics* (Brooke and Hirst, 2013a).

namely on deriving concreteness at the lexical level. The MRC psycholinguistic database (Coltheart, 1980) has manual annotations for degree of concreteness, though their definitions are somewhat different than ours.[21] Changizi (2008) based their notion of concrete and abstract on the depth in the WordNet hierarchy, though, as Turney et al. (2011) point out, this is really more a measure of specificity than concreteness. Turney et al. (2011) use a version of the LSA method that we have already discussed in Section 3.3 and applied to formality in Section 3.4; they use the MRC database as a seed set. Having built a lexicon, they use it to differentiate concrete and abstract senses of words.

3.6.2 Model

Our main model is an adaption of the popular latent Dirichlet allocation (LDA) topic model (Blei et al., 2003), with each of the 6 styles corresponding to a topic. We also test the correlated topic model (CTM) for these purposes. We discussed the specifics of these models earlier in Section 3.2: here, β_z corresponds to corresponding to the probability of a topic z generating any given word in the vocabulary.[22] For both LDA and CTM we use the original variational Bayes implementation of Blei. Variational Bayes (VB) works by approximating the true posterior with a simpler distribution, minimizing the Kullback-Leibler divergence between the two through iterative updates of specially-introduced free variables. The mathematical and algorithmic details are omitted here; see Blei et al. (2003; 2007). Our early investigations used an online, batch version of LDA (Hoffman et al., 2010), which is more appropriate for large corpora because it requires only a single iteration over the dataset. We discovered, however, that batch models were markedly inferior to more traditional models for our purposes because the influence of the initial model diminishes too quickly; here, we need particular topics in the

[21] For example, the dictionary considers many function words to be highly abstract, whereas we would consider them as neutral or even concrete.

[22] Some versions of LDA smooth this distribution using a Dirichlet prior; here, though, we use the original formulation from Blei (2003), which does not.

model to correspond to particular styles, and we accomplish this by seeding the model with known instances of each style. Specifically, our initial β consists of distributions where the entire probability mass is divided amongst the seeds for each corresponding topic, and a full iteration over the corpus occurs before β is updated. Typically, LDA iterates over the corpus until a convergence requirement is met, but in this case this is neither practical (due to the size of our corpus) nor necessarily desirable; the diminishing effects of the initial seeding means that the model might not stabilize, in terms of its likelihood, until after it has shifted away from our desired stylistic dimensions towards some other variation in the data. Therefore, we treat the optimal number of iterations as a variable to investigate.

The model is trained on a 1 million text portion of the ICWSM Spinn3r dataset 2009 (Burton et al., 2009). Since our method relies on co-occurrence, we followed our earlier work in using only texts with at least 100 different word types. All words were tokenized and converted to lower-case, with no further lemmatization. Following Hoffman et al. (2010), we initialized the α of our models to $1/k$ where k is the number of topics. Otherwise we used the default settings; when they overlap they were identical for the LDA and CTM models.

3.6.3 Lexicon induction

Our primary evaluation is based on the stylistic induction of held-out seed words. Our definition of the words that belong in each style is given below.

Colloquial Words which are used primarily in very informal contexts, for instance slang words and internet abbreviations.

Literary Words which you would expect to see primarily in literature; these words often feel old-fashioned or flowery.

Concrete Words which refer to events, objects, or properties of objects in the physical world that you would be able to see, hear, smell, or touch.

Abstract Words which refer to something that requires major psychological or cultural knowledge to grasp; complex ideas which can't purely be defined in physical terms.

Subjective Words which are strongly emotional or reflect a personal opinion.

Objective Words which are emotionally distant, explicitly avoiding any personal opinion, instead projecting a sense of disinterested authority.

The words were collected from various sources by the author, a native speaker of English with significant experience in English linguistics. Included words had to be clear, extreme members of their stylistic category, with little or no ambiguity with respect to their style. The colloquial seeds consist of English slang terms and acronyms, e.g. *cuz*, *gig*, *asshole*, *lol*. The literary seeds were primarily drawn from web sites which explain difficult language in texts such as the Bible and *Lord of the Rings*; examples include *behold*, *resplendent*, *amiss*, and *thine*. The concrete seeds all denote objects and actions strongly rooted in the physical world, e.g. *shove* and *lamppost*, while the abstract seeds all involve concepts which require significant human psychological or cultural knowledge to grasp, for instance *patriotism* and *nonchalant*. For our subjective seeds, we used an edited list of strongly positive and negative terms from a manually-constructed sentiment lexicon (Taboada et al., 2011), e.g. *gorgeous* and *depraved*, and for our objective set we selected words from sets of near-synonyms where one was clearly an emotionally-distant alternative, e.g. *residence* (for *home*), *jocular* (for *funny*) and *communicable* (for *contagious*). We filtered initial lists to 150 of each type, removing words which did not appear in the corpus or which occurred in multiple lists. For evaluation we used stratified 3-fold crossvalidation, averaged over 5 different (3-way) splits of the seeds, with the same

splits used for all evaluated conditions.

Given two sets of opposing seeds, we follow our earlier work on formality in evaluating our performance in terms of the number of pairings of seeds from each set which have the expected stylistic relationship relative to each other (the guessing baseline is 0.5). Given a word w and two opposing styles (topics) p and n, we place w on the PN dimension according to the β of our trained model as follows:

$$PN_w = \frac{\beta_{pw} - \beta_{nw}}{\beta_{pw} + \beta_{nw}}$$

The normalization is important because otherwise more-common words would tend to have higher PN's, when in fact the opposite is true (rare words tend to be more stylistically prominent). We then calculate pairwise accuracy as the percentage of pairs $\langle w_p, w_n \rangle$ ($w_p \in P_{seeds}$ and $w_n \in N_{seeds}$) where $PN_{w_p} > PN_{w_n}$. However, this metric does not address the case where the degree of a word in one stylistic dimension is overestimated because of its status on a parallel dimension. Two more-holistic alternatives are total accuracy, the percentage of seeds for which the highest β_{tw} is the topic t for which w is a seed (guessing baseline is 0.17), and the average rank of the correct t as ordered by β_{tw} (in the range 1–6, guessing baseline is 3.5); the latter is more forgiving of near misses.

We tested a few options which involved straightforward modifications to model training. Standard LDA produces all tokens in the document, but when dealing with style rather than topic, the number of times a word appears is much less relevant. Our binary model assumes an LDA that generates types, not tokens.[23] A key comparison here is with a combined LDA

[23] At the theoretical level, this move is admittedly problematic, since our LDA model is thus being trained under the assumption that texts with multiple instances of the same type can be generated, when of course such texts cannot by definition exist. We might address this by moving to Bayesian models with very different generative assumptions, e.g. the spherical topic model (Reisinger et al., 2010), but these methods involve a significant increase of computational complexity and we believe that on a practical level there are no real negatives associated with directly using a binary representation as input to LDA; in fact, we are avoiding what appears to be a much more serious problem, burstiness (Doyle and Elkan, 2009), i.e. the fact that traditional LDA is influenced too much by multiple instances of the same word.

Table 3.6: Model performance in lexical induction of seeds. Bold indicates best in column.

Model	Pairwise Accuracy (%)				Total	Avg.
	Lit/Col	**Abs/Con**	**Obj/Sub**	**All**	Acc. (%)	Rank
guessing baseline	50.0	50.0	50.0	50.0	16.6.	3.50
basic LDA (iter 2)	94.3	98.8	93.0	95.4	55.0	1.79
binary LDA (iter 2)	96.2	98.9	**93.5**	**96.2**	**57.7**	**1.74**
combo binary LDA (iter 1)	95.4	**99.2**	93.3	96.0	53.1	1.86
binary CTM (iter 1)	**96.3**	99.0	89.6	95.0	53.0	1.87

model (*combo*), an amalgamation of three independently trained 2-topic models, one for each dimension; this tests our key hypothesis that training dimensions of style together is beneficial. Finally, we test against the correlated topic model, which offers an explicit representation of style correlation, but which has done poorly with respect to interpretability, despite offering better likelihood (Chang et al., 2009).

The results of the lexicon induction evaluation are in Table 3.6. Since the number of optimal iterations varies, we report the result from the best of the first five iterations, as measured by total accuracy; the best iteration is shown in parentheses. In general, all the results are high enough—we are reliably above 90% for the pairwise task, and above 50% for the 6-way task—for us to conclude with some confidence that our model is capturing a significant amount of stylistic variation. As predicted, using words as boolean features had a net positive gain, consistent across all of our metrics, though this effect was not as marked as we have seen previously. The model with independent training of each dimension (*combo*) did noticeably worse, supporting our conclusion that a multidimensional approach is warranted here. Particularly striking is the much larger drop in overall accuracy as compared to pairwise accuracy, which suggests that the combo model is capturing the general trends but not distinguishing correlated styles as well. However, the most complex model, the CTM, actually does slightly worse than the combo, which was contrary to our expectations but nonetheless consistent with previous work on the interpretability of topic models. The performance of the full LDA mod-

els benefited from a second iteration, but this was not true of combo LDA or CTM, and the performance of all models dropped after the second iteration.

An analysis of individual errors reveals, unsurprisingly, that most of the errors occur across styles on the same pole; by far the largest single common misclassification is objective words to abstract. Of the words that consistently show this misclassification across the runs, many of them, e.g. *animate*, *aperture*, *encircle*, and *constrain* are clearly errors (if anything, these words tend towards concreteness), but in other cases the word in question is arguably also fairly abstract, e.g. *categorize* and *predominant*, and might not be labeled an error at all. Other signs that our model might be doing better than our total accuracy metric gives it credit for: many of the subjective words that are consistently mislabeled as literary have an exaggerated, literary feel, e.g. *jubilant*, *grievous*, and *malevolent*.

3.6.4 Text-level analysis

Our secondary analysis involved evaluating the θ's of our best configuration (based on average pairwise and total accuracy) on other texts in a mixed genre corpus, to see whether our dimensions correspond to our intuitions about individual genres. After training, we carried out inference on the BNC corpus, averaging the resulting θ's to see which styles are associated with which genres. Appearances of the seed terms for each model were disregarded during this process; only the induced part of the lexicon was used. The average differences relative to the mean across the various stylistic dimensions (as measured by the probabilities in θ) are given for a selection of genres in Table 3.7.

The most obvious pattern in Table 3.7 is the dominance of the medium: all written genres are positive for our styles on the 'cultural' pole and negative for styles on the 'situational' pole and the opposite is true for spoken genres. The magnitude of this effect is more difficult to interpret: though it is clear why fiction should sit on the boundary (since it contains spoken

Table 3.7: Average differences from corpus mean of LDA-derived stylistic dimension probabilities for various genres in the BNC, in hundredths.

Genre	Styles					
	Literary	**Abstract**	**Objective**	**Colloquial**	**Concrete**	**Subjective**
News	+0.67	+0.50	+0.43	−0.31	−0.72	−0.57
Religious texts	+0.38	+0.38	+0.28	−0.27	−0.44	−0.32
Academic	+0.18	+0.29	+0.26	−0.20	−0.36	−0.18
Fiction	+0.31	+0.09	+0.02	−0.05	−0.12	−0.25
Meeting	−0.61	−0.54	−0.42	+0.35	+0.69	+0.55
Courtroom	−0.63	−0.53	−0.41	+0.32	+0.69	+0.57
Conversation	−0.56	−0.63	−0.54	+0.43	+0.80	+0.50

dialogue), the appearance of news at the written extreme is odd, though it might be due to the fact that news blogs are the most prevalent formal genre in the training corpus.

However, if we ignore magnitude and focus on the relative ratios of the stylistic differences for styles on the same pole, we can identify some individual stylistic effects among genres within the same medium. Relative to the other written genres, for instance, fiction is, sensibly, more literary and much less objective, while academic texts are much more abstract and objective; for the other two written genres, the spread is more even, though relative to religious texts, news is more objective. At the situational pole, fiction also stands out, being much more colloquial and concrete than other written genres. Predictably, if we consider again the ratios across styles, conversation is the most colloquial genre here, though the difference is subtle.

We carried out a correlation analysis of the LDA-reduced styles of all texts in the BNC and, consistent with the genre results in Table 3.7, found a strong positive correlation for all styles on the same main pole, averaging 0.83. The average negative correlation between opposing poles is even higher, -0.88. This supports the Leckie-Tarry formulation of correlated styles. Notably, the independence assumptions of the LDA model did not prevent strong correlations from forming between these distinct yet clearly interrelated dimensions; if anything, the correlations are stronger than we would have predicted, and may require more targeted effort to distinguish.

3.7 Hybrid models for multi-dimensional style

3.7.1 Introduction

In this section, we expand our investigation of the multi-dimensional stylistic space introduced in the preceding section.[24] Key differences include an annotation study which will allow for a more reliable pairwise evaluation metric, the inclusion of LSA and PMI as options for extracting stylistic information from corpora, and the use of other methods post-extraction to further distinguish the styles. Note that unlike all preceding work, we no longer make explicit assumptions about which styles are opposing.

3.7.2 Word annotation

We continue to use the 900 word set, with 150 words hand-picked to represent each style. However, relying on a single annotator as we have done thus far is problematic, and a more serious issue with our original seed sets is that many of the seeds belong on multiple lists, reflecting the fact that stylistic correlations occur at the lexical level. This interferes with evaluation, since we need to to be fairly certain not only which seeds are in a category, but which aren't. Therefore, we carried out a full annotation study with 5 annotators, asking each annotator to tag all 900 words for each of the 6 styles according to guidelines we prepared. I included myself as an annotator (this annotation was carried out prior to all the others, and not further modified), but the other four were unfamiliar with the project; all were native English speakers with at least an undergraduate degree, and all reported reading a variety of text genres for work and/or pleasure. We provided written guidelines explaining each style in detail, and asked annotators to make judgments based on what they felt to be the most common sense. Communication among annotators was restricted during the process, but we allowed access

[24] The work presented in this section is based on "Hybrid models for lexical acquisition of correlated styles" by Julian Brooke and Graeme Hirst, published in *Proceedings of the 6th International Joint Conference on Natural Language Processing*.

Table 3.8: Fleiss's kappa for 5-way annotation, by style.

Style	Kappa
Literary	0.61
Abstract	0.37
Objective	0.55
Colloquial	0.85
Concrete	0.67
Subjective	0.63
Average	0.61

to other resources (e.g. the internet) and answered general questions about the guidelines that came up during the process. A few annotators had obviously skewed numbers for certain styles relative to other annotators due to misinterpretation of the guidelines, and we provided non-specific feedback for revision in these cases. The Fleiss's kappa (Fleiss, 1971) values for our 5-way annotation study are presented in Table 3.8.

The kappa values in Table 3.8 indicate agreement well above chance, but several of the dimensions (and the average) are below the 0.67 standard for reliable annotation (Artstein and Poesio, 2008), and only one (colloquial) reaches the higher 0.8 standard. This suggests that there is a sizable subjective aspect to these judgments and we should be somewhat skeptical of the judgment of any particular annotator. However, we had forced our annotators to make a boolean choice for each style, which may be somewhat inappropriate for somewhat non-discrete phenomena like stylistic variation. Taboada et al. (2011), when validating their fine-grained manual polarity lexicon (which included annotation of both polarity and strength), demonstrated that Mechanical Turk worker disagreement on a boolean task seemed to correspond fairly well to ranges on a scale: there was agreement at the extremes of polarity, but increasing disagreement towards the middle.

With this in mind, we used our initial annotations to create a new annotation task for two of our external annotators: the goal was to investigate whether annotators can identify rela-

tive differences in degree suggested by either agreement or disagreement with their choices by other annotators. First, we extracted minority opinions, defined here as word/style combinations where the annotator agreed with exactly one other annotator and disagreed with the three others, and consensus opinions, defined as those where all the annotators agreed. We randomly paired each minority opinion word/style with a consensus opinion; for both opinions, the annotator in question had made the same judgment (both yes, or both no), but some of the other annotators had made different choices. We then asked our annotators (who were unaware of the exact nature of the experiment) to pick, among two words they had tagged the same in the first round, the word which had 'more' of the relevant stylistic quality.

In the negative case (where the annotator had originally marked both as not having the style), the results are stark: in 97% of the cases, the annotator picked the minority opinion (i.e. the word which some other annotators had marked yes), suggesting that the annotator could identify the stylistic tendencies of the (mixed-agreement) word, but had nonetheless excluded it, probably because there were much clearer examples of this style and other styles which could be more clearly applied to the word. In the positive case, the annotators preferred the word with group consensus 82.7% of the time, which is indeed the pattern we would predict if the minority opinion is less extreme; the positive case is more subtle than the negative case, where many of the words used for comparison very clearly do not belong to the relevant style. These results are consistent with the idea that disagreement is a rough indicator of degree, and that not all disagreement should be dismissed as noise or some other failure of annotation. Of course, this also indicates that relative or continous (e.g. Likert scale) judgments might be preferable to boolean ones, but in this case boolean annotation is far more practical, and indeed desirable for both model creation and evaluation.

For our final seed set, our positive annotations include all word/style combinations where a majority of annotators marked yes, whereas our negative annotations include only terms where

Table 3.9: Number of seeds, by style.

Style	Positive	Negative
Literary	132	660
Abstract	107	599
Objective	245	495
Colloquial	163	684
Concrete	190	572
Subjective	258	487

there was complete consensus; words where only 1 or 2 annotators marked yes were removed from consideration as seeds (for that particular style). The summary of the counts for main seed set are presented in Table 3.9.

3.7.3 Methods

Our method for stylistic lexicon acquisition breaks down into three steps. The first is to apply one of several methods which leverages co-occurrence in a large corpus to derive, for each word, a raw score for each style. We then take that raw score and normalize it; the resulting number can be used directly to compare words relevant to a style. Finally, we consider the vector formed by these normalized style scores, and apply other methods which further refine this vector, implicitly taking into account the correlations among styles. The elements of the refined vector correspond to the degree of each style, so if we apply this method for all words in our vocabulary we create a full-coverage lexicon.

3.7.4 Corpus analysis

For all the methods in this section, we use the same corpus, the ICWSM Spinn3r 2009 dataset (Burton et al., 2009), which has been used successfully in related work (Brooke et al., 2010a). Social media corpora are particularly appropriate for research on style, since they contain a variety of registers. Here, we include all 2.46 million texts in the Tier 1 portion which contained

at least 100 word types. Hapax legomena were excluded, since they could not possibly offer any co-occurrence information, but otherwise we did not filter or lemmatize words: our full vocabulary is 1.95 million words.

Our simplest method uses pointwise mutual information (PMI) (Church and Hanks, 1990), a popular metric for measuring the association between words. Since standard PMI has a lower bound of $-\infty$ when the joint probability is 0 (a common occurrence since many of our words are relatively rare), we actually use a normalized version, NPMI, which has an upper bound of 1 and a lower bound of -1.

$$NPMI(x,y) = \left(\log \frac{p(x,y)}{p(x)p(y)}\right)\left(\frac{1}{\log p(x,y)}\right)$$

Following our earlier work, here and elsewhere we do not use the term frequency within a document; instead the probabilities are calculated using the number of documents where the word or words appear divided by the total number of documents. The raw score r_{ij} for style i of word w_j is simply the sum of its NPMI with the associated set of seeds S_i:

$$r_{ij} = \sum_{s \in S_i} NPMI(w_j, s)$$

Our second method is LSA. Assuming $\mathbf{v}_w$ denotes the resulting k-dimensional vector for word w, we calculate r_{ij} as:

$$r_{ij} = \sum_{s \in S_i} \cos(\theta(\mathbf{v}_{w_j}, \mathbf{v}_s))$$

Our third method uses latent Dirichlet allocation (Blei et al., 2003), as outlined in the preceding section. There, we found that two iterations were preferred, and we continue that here, along with a binary feature representation. For the LDA method, r_{ij} corresponds directly to β_{ij} of the resulting model which is just the probability of topic (style) i generating w_j.

3.7.5 Normalization

The raw numbers derived from corpus analysis methods discussed above cannot be used directly as indicators of style: the frequencies of both the seeds and the words being predicted have significant effect on the relative and absolute magnitudes of each style for all our methods, and performance using just these numbers is near chance. However, in two steps we can normalize these numbers to a form where the magnitude does directly reflect degree of a style. Again, r_{ij} refers to the raw score for style i and word j from some corpus analysis method. First, we take steps to ensure that r_{ij} is nonnegative. For LDA this is unnecessary (since r_{ij} is based on a probability distribution), but for NPMI and LSA it is needed, since both involve summing over items which vary between -1 and 1. We can ensure that these are positive by adding a constant equal to the number of seeds. Next, we convert the result to a style 'distribution' for each word:

$$r'_{ij} = \frac{r_{ij} + |S_i|}{\sum_{k=1}^{6} r_{kj} + |S_k|}$$

The result is still not useful, since frequency (and count) of seeds clearly still has an effect. To focus on the differences between words, we subtract the means for each style and divide by the standard deviation

$$b_{ij} = \frac{r'_{ij} - \overline{r'_i}}{\sigma_{r'_i}}$$

to reach b_{ij}, the base for the 'style space' methods in the next subsection.

3.7.6 Style vector optimization

Given a vector that represents the styles for a given word, we wish to refine the vector to improve performance on relative judgments for individual styles. Here, we test two options:

the first transforms the stylistic vectors into k-Nearest Neighbor (kNN) graphs, where we can apply label propagation. The second option treats the vector as a set of features for supervised linear regression, one for each style, using a specialized loss function. Both methods rely on having a style vector representation of not only our target words, but also our seed (training) words. For LSA and NPMI, we used leave-one-out crossvalidation to create these vectors; for LDA, however, it was impractical to do a full run of the model for each word, and so we used 10-fold crossvalidation instead.

A vector-space representation offers a number of obvious similarity functions for building a kNN graph: we test two here, inverse Euclidean distance (L2) and cosine similarity (cos). A more difficult problem is the choice of k (for kNN k): here, we estimate a good k from the training set. Since the training set and dimensionality of the data is (now) fairly small, we simply test on all possible intervals of 5, and choose the best (often near 50, though we saw values as low as 10 and as high as 90) using our pairwise evaluation metric discussed in the next section. Since our label propagation method works independently for each style, we can choose a different k for each.

For label propagation, we use the simple one-step propagation function from Kang et al. (2006). Here, K is our similarity function (which returns zero if seed s is not one of the k nearest neighbors), and z_{ij} is the resulting confidence score, which we use as our new estimate for the style:

$$z_{ij} = \sum_{w_s \in S_i} K(w_j, w_s)$$

Obviously, the main work here is done by the similarity function, which implicitly includes information from other stylistic dimensions by preferring words which are close not just on the relevant dimension, but in the stylistic space as a whole. There are of course more sophisticated, multi-step approaches to label propagation, e.g. the one used by Rao and Ravichandran (2009),

but a single-step approach has clear advantages in light of our large vocabulary and dense graph; we leave exploration of whether unlabeled words can help further to future work. We did test the one-step correlated label propagation method proposed by Kang et al. but found it was ineffective, probably because it increases the effects of correlation, which is actually counter to our needs.

The information provided by label propagation is distinct enough that it can be successfully combined with the original (base) vector. As with k for kNN, we estimated a good weighting for this combination using the training data, testing at 0.01 intervals. Since we noted some interdependence, we combined this step with the selection of (kNN) k. Again, this ratio can be different for each style.

Our second vector optimization technique is an adaption of supervised linear regression. Linear regression usually involves minimizing squared distance of the output of the model from the training set, assuming there are known values of expected output. In this case, however, we don't have reliable values for specific degrees of a style. We proceed by replacing the least-squared loss function with a loss function based on our evaluation metric:

$$L(\theta) = \sum_{w_j \in S_{i,p}} \sum_{w_m \in S_{i,n}} I(h_\theta(b_{ij}) < h_\theta(b_{im}))$$

Here, $S_{i,p}$ and $S_{i,n}$ refer to the positive and negative examples of style i, respectively, h_θ is the linear regression function, and I is an indicator function equal to 1 if the statement is true, and 0 otherwise.

Using such a loss function discourages standard approaches to linear regression, but in this context (a small feature space and training set), it is reasonably practical to search the space exhaustively for weights which provide a (near-)optimal result (on the training data). Starting with full weight (1) on the feature corresponding to the dimension being derived and 0 on all others, we search the range -1 to 1 at 0.001 intervals for the other dimensions, proceeding in

order based on the greatest difference across positive and negative examples of each style. We found that one such iteration across each element of the vector was sufficient, resulting in a stable model. This method can be applied on the initial vector, or on a vector that has already been refined by some other method, i.e. the output of label propagation.

3.7.7 Evaluation

Our evaluation is based on the pairwise comparison of words which are known (from our annotation) to differ relevant to a certain style. Accuracy for a test set S_i (of a style i) is defined as the number of instances where the expected inequality exists between a pair of opposing words, divided by the total number of such pairings:

$$Accuracy(S_i) = \frac{\sum_{w_j \in S_{i,p}} \sum_{w_m \in S_{i,n}} I(z_{ij} > z_{im})}{|S_{i,p}| \cdot |S_{i,n}|}$$

Here z can refer to any of the metrics for style discussed in the previous section. The major advantage of this definition of accuracy is that it does not require an arbitrary cutoff point, but 100% accuracy nonetheless indicates that the two sets are perfectly separable. Also, it does not assume anything about the degree of difference between two words, e.g. that more is better, since for any given pair of words we cannot be certain what an ideal difference would be.

We evaluate using 3-fold crossvalidation, using the original 150-per-style annotation of our 900 words for the purposes of stratifying the data, which allows for balanced sets of 600 for training and 300 for testing. All seeding, training, and evaluation use the majority annotation of the 5 annotators as discussed earlier. Since the initial splits add a significant random factor, all results here are averaged over 5 runs, with the same 5 runs (i.e. same splits) used for all evaluated conditions.

Table 3.10 shows a comparison of the performance of various models, organized by the method of corpus analysis. First, we note that most of these numbers are quite high, almost

Table 3.10: Model performance in lexical induction of seeds, % pairwise accuracy. LP = label propagation, cos = cosine similarity, L2 = inverse Euclidean distance, LR = linear regression. Bold is best in column.

Model	By Style						Average
	Lit.	Abs.	Obj.	Coll.	Conc.	Subj.	
guessing baseline	50.0	50.0	50.0	50.0	50.0	50.0	50.0
NPMI							
base (Normalized)	68.4	91.2	94.4	95.6	73.4	77.1	83.0
LP-cos	90.1	91.5	95.1	94.4	90.0	80.0	90.2
LP-L2	88.2	88.9	94.1	94.1	89.4	76.6	88.5
base+LP-cos	90.2	92.8	95.6	96.0	90.6	80.9	91.0
base+LR	89.8	93.6	94.2	96.5	85.5	79.7	89.9
base+LP-cos, LR	90.2	93.6	95.5	95.9	90.5	81.0	91.1
LDA							
base	67.3	93.3	96.5	96.2	93.2	83.5	88.3
LP-cos	86.0	92.9	96.0	93.6	94.8	86.5	91.6
LP-L2	78.1	91.1	95.0	92.5	94.2	83.2	89.0
base+LP-cos	86.4	93.5	96.6	96.3	95.5	86.7	92.5
base, LR	84.3	93.9	96.5	96.4	94.7	85.7	91.8
base+LP-cos, LR	87.2	93.9	96.5	96.3	**95.8**	**87.0**	92.8
LSA							
k=20, base	89.1	93.5	95.6	94.4	90.8	76.0	89.9
k=500, base	91.2	93.7	96.5	96.5	93.7	83.5	92.6
k=500, LP-cos	92.4	91.7	96.0	96.8	94.3	85.2	92.8
k=500, LP-L2	92.1	92.1	96.5	96.5	94.3	85.0	92.8
k=500, base+LP-cos	92.5	93.6	96.8	97.5	94.8	85.9	93.5
k=500, base, LR	**92.7**	**94.0**	**97.2**	97.2	94.9	86.5	**93.7**
k=500, base+LP-cos, LR	**92.7**	93.8	97.0	**97.7**	94.9	86.4	**93.7**

all are above 80% and most are above 90%. It is worth mentioning that if only direct opposites are considered (e.g. colloqiual versus literary, concrete versus abstract), most dimensions reach results above 99%; our multi-style evaluation here offers a more realistic view. Among individual styles, colloquial words seem the most distinct, which is consistent with the results of human annotation. Acquisition of subjectivity, on the other hand, is strikingly more difficult than the other styles.

Based only on average accuracy, we could conclude that LSA $>$ LDA $>$ NPMI with respect to extracting relevant stylistic information from the corpus. That NPMI is the worst perform-

ing method is not surprising, since it relies only on direct co-occurrence between seeds and test words, and is not able to take advantage of larger patterns in the data; we would expect similar results for other simple relatedness measures. Though LSA is better overall, the distinction between LSA and LDA is more subtle, since in fact LDA is the higher performing model for 2 of the 6 styles, and its poorer overall performance can almost entirely be attributed to a rather dismal showing for literary words, worse than NPMI. This is interesting because subjective and concrete words, where LDA does well, are the most common in the corpus, whereas literary words are consistently the least common. We posit, based on this and our earlier research focused on the LDA method, that successful low-dimensional seeded LDA requires styles (topics) that are reasonably well-represented in the corpus; when that condition is met, LDA will likely do better than LSA because it will distinguish rather than collapse correlated styles. LSA, on the other hand, is robust against the scarcity problem because it requires only that a set of words have a reasonably distinct k-dimensional profile to form a coherent style.

Based on the results in Table 3.10, we can conclude decisively that both of our optimization techniques are effective. The effects are particularly marked for NPMI, but are reasonably consistent across all three corpus analysis techniques and the various individual styles. With regards to the similarity function in label propagation, we found that cosine similarity, a less common choice for building graphs, was generally as good as, and often better than, Euclidean distance. The vector resulting from label propagation also consistently benefited from being combined with the base vector, the result being better than either alone. It is not entirely clear which of the two optimization methods is to be preferred (their effects seem roughly similar), though linear regression seems to have the edge when using LSA. Combining the two methods seems a good strategy, particularly for LDA.

The LSA results presented here mostly use $k = 500$, a fairly standard choice. However, we tested other values, in particular extremely low values ($k = 20$) to see if we could confirm

the our earlier supposition that much stylistic information is primarily contained with the first few dimensions of LSA. Our results suggest that the basic supposition is valid, since the difference between the two conditions for most dimensions is not large, but the identification of subjectivity (not considered earlier) does seem to benefit greatly from a higher-dimensional vector.

3.7.8 Qualitative analysis

To investigate further the successes and failures of our method, we carried out two qualitative examinations of the output of our model. First, we looked at those words within our annotated set of words which consistently caused the most errors across the various splits and runs. Second, we ran a high-performing LSA model built from the entire seed set on a subset of our vocabulary (we excluded words of document frequency less than 100), creating lexicons for each style; we manually inspected non-seed words that were ranked highest on each dimension.

The clearest result from the inspection of the seed output was that many of the false negatives involve words that are strong in some other dimension, typically on the other side of the oral/literate divide. For example, the most difficult-to-identify literary and abstract terms are strongly subjective (e.g. *loathe* and *obscene*), while the most difficult objective word, *translucent*, is very concrete. The most difficult concrete words are literary (*yoke*, *raiment*) or objective (*conflagration*), and the most difficult subjective words are also somewhat objective (*eminent*) or abstract (*autocratic*). Interestingly, a manual inspection of the weights for linear regression suggests that our optimization is correcting for just this kind of situation: we generally see negative weights on (what we would predict to be) positively correlated styles, and positive weights on negatively correlated styles. However, in certain cases where one style has a much larger role in determining the co-occurrence pattern in the corpus, this correction may

be insufficient.

Most of the false positives, by contrast, involve overextension of each category in predictable ways. For example, our highest ranking literary words from the general vocabulary were mostly very good, but contained a few words that are obvious overgeneralizations into biblical and fantasy texts, e.g. *locust* and *sorcerers*, while among the objective words there were a number of academia-relevant words that are really more abstract than objective, e.g. *coauthors* and *peer-review*. Our derived colloquial words contained many (sometimes purposeful) misspellings (*wayy*, *annnnd*) which we could argue are genuinely colloquial; less clear are the many lower-case celebrity names (e.g. *miley*), but the fact that the bloggers used lower case does make them non-standard. Consistent with our qualitative results, subjective was the most problematic style in the general vocabulary: though there were many good subjective words, there were a lot of other words which suggest topics that people tend to express opinions about, e.g. *sitcoms*, *entertainer*, or *flick*; movie-related words are particularly common, which might be a reflection of the lexicon we took our subjective seeds from.

3.8 Supervised sociolinguistic variable identification

3.8.1 Introduction

The project discussed in this section is a departure from the lexicon creation we have dealt with so far: its intention is not to create a high-coverage lexicon for the benefit of downstream tasks, but rather to identify a set of features for sociolinguistic study.[25] In this case, we use a transcribed spoken corpus which is already tagged for social factors of interest, so we can apply a simple supervised feature selection metric to isolate particular words and patterns that are of interest to a sociolinguist. A significant manual component is required.

[25] Aspects of this work were originally presented as "Hunting the linguistic variable: Using computational techniques for data exploration and analysis" by Julian Brooke and Sali Tagliamonte, at the Georgetown University Round Table on Languages and Linguistics 2012 (Brooke and Tagliamonte, 2012).

3.8.2 Corpus

The Toronto Corpus (Tagliamonte, 2006) includes 200 transcribed interviews with individuals who have lived their entire life in the city of Toronto; it was carefully collected to reflect a wide range of social backgrounds, including annotations for age, gender, ethnicity, education, and social class. It was not, initially, machine-readable, due to some inconsistencies in the formatting; I was, however, able to fix these inconsistencies and create a machine-readable version; at the same time, I tokenized and tagged it using the TreeTagger (Schmid, 1995). The corpus contains the text of both interviewee and interviewer, but for the work discussed here, we only consider the text of the interviewee, from which I also removed all markers in the annotation (e.g. laughter, coughing).

3.8.3 Method

Given the tagged Toronto corpus, I extracted all mixed word and POS *n*-grams (unigrams, bigrams, and trigrams) that appeared at least 5 different interviews. By mixed, I mean that for each slot in the *n*-gram, I considered both the POS and word as a legitimate option, so for bigram *the book* there are actually 4 mixed *n*-grams, namely *the book*, *the NN*, *DT book*, and *DT NN* The idea behind using a mixture of words and POS was to be able to narrow in on particular lexicogrammatical patterns, for instance the use of *so* as an intensifier (appearing always before an adjective) rather than as a connective.

For some of the social factors (e.g. ethnicity) there were simply too many categories and too few examples of each, so I focused on 4 which could be simplified to two classes: gender (male or female), education (post-secondary or no post-secondary), social class (white collar or blue collar), and age (over 30 or under 30). For these classes, I used the WEKA machine learning suite (Witten and Frank, 2005) to rank all the features according to their information gain, supposing we were using the frequency of the feature to predict the relevant social factor.

Information gain represents the decrease in entropy (i.e. uncertainty) relevant to a classification when a particular feature is known as compared to when it is not known. If x is a random variable, entropy is defined as:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log x_i$$

And IG for some variable y (our feature) which may predict x is:

$$IG(y) = H(x) - H(x|y)$$

Information gain is used, perhaps most notably, as the metric for choosing the next feature to split a decision tree. We are not, however, interested in a optimal classification model here, just the features that might be used to build one.

3.8.4 Analysis

Next, we manually inspected the first several hundred best features for each category. Many of the most predictive words and phrases were variations on well-established markers in sociolinguistics, for example the prevalence of *was like* and intensifier *so* among younger women. Others were indicators that were predictable but uninteresting from the perspective of sociolinguistics (e.g. *my wife* indicates an older male, talk of *the office* indicated a white-collar worker). Some were completely opaque, possibly the result of random factors: for instance, determiners before or after disfluencies were a fairly strong indicator of gender. However, there were numerous examples that fell between obvious and inscrutable; for some of these cases, I paired them with synonyms and ran a variationist analysis to confirm that they were statistically significant. The word *supper*, for example, is essentially never used by the young (or even middle-aged), whereas the word *weird* is a distinctly young way of expressing *strange*. Another word for *strange*, *odd*, is preferred by the older and educated. I found that *a number*

of is a white-collar way to say *a lot*, whereas *a bunch of* is more blue-collar; both, however, are preferred by men. There are many more examples for all the factors, enough to fill a small lexicon. However, as it stands now, human intervention is required at some stage; otherwise a lot of duplicates, garbage, and topic-relevant words would be included. Nevertheless, the use of feature selection revealed variables that would not have been noticed in a manual inspection and has led to a larger exploration of adjectives, in particular (Tagliamonte and Brooke, submitted).

Chapter 4

Stylistic Tasks

The construction of stylistic lexicons using automated methods is not generally an end in itself. In this chapter, I first identify several major areas of NLP which are examples of potential extrinsic evaluations for the usefulness of such a resource. My own contributions in this regard form the rest of the chapter, a summary of these is given in Table 4.1.

4.1 Survey

4.1.1 Text classification

Style played an important role in the early development of automatic text classification: the feature sets of early approaches to genre classification (Karlgren and Cutting, 1994; Kessler et al., 1997) were based on (easily derived) surface features from Biber (1988), including counts of major parts of speech, words per sentence, and long words. Kessler et al. (1997), in particular, use stylistic lexical features such as words with Latinate roots. Other genre-focused work that is relevant to style includes that of Finn et al. (2002), who show that *part-of-speech* (POS) ratios are more effective features than a *bag of words* (BOW) for distinguishing between objective and subjective texts across genres. In most of this work, simple textual features or a kitchen-sink BOW approach are a substitute for a more nuanced understanding of the style of individual lexical items.

Table 4.1: Overview of contributions in Chapter 4, including tasks investigated, methods used, and conclusions reached.

- Section 4.2

  Task Register differentiation

  Methods Multidimensional analysis, LSA, qualitative analysis, transfer across spaces

  Conclusions Dimensionality reduction technique used is unimportant, bag-of-words offers similar or even better results than MD features, robust across corpus/feature spaces

- Section 4.3

  Task Sociolinguistic variable differentiation

  Methods Formality score (Section 3.4)

  Conclusions Formality distinguishes education, age, and gender of interviewer

- Section 4.4

  Task Clipping prediction

  Methods Formality score (Section 3.4), LSA vector classification, crowdsourced annotation

  Conclusions Formality score as good as full LSA vector, low k preferred

- Section 4.5

  Task Segmentation of poetry

  Methods Vector space methods, stylistic change curve, extrinsic lexical features

  Conclusions Extrinsic features useful, formality and low k LSA best, real task harder than artificial

- Section 4.6

  Task Clustering of voices in poetry

  Methods Vector space methods, k-means clustering, cluster seeding

  Conclusions Basically works, but overall performance strongly dependent on segmentation, seeding helps quite a bit

- Section 4.7

  Task Intrinsic plagiarism detection

  Methods Vector space methods, clustering, effect of expected difference

  Conclusions Taking account of expected differences from span size helps, but task evaluation is poor

- Section 4.8

  Task Distinguishing sub-genres using style

  Methods 6-style lexicons (Section 3.7)

  Conclusions Mostly predicted distribution, surprises inform literary study

Other work in text classification has explicitly focused on classification of stylistic differences. Argamon et al. (1998) highlight POS trigrams and function words as being indicative of style as compared to topic; they use them to distinguish effectively among articles from different newspapers. Work by Koppel (Koppel et al., 2005) identifies stylistic features that are useful for classifying the native language of non-native writers (See Chapter 5). Emigh and Herring (2005) compare the formality of online collaborative encyclopedias using the contextuality measures of Heylighen and Dewaele (2002). Peterson et al. (2011) identify formal and informal e-mail messages in the Enron corpus in order to test sociolinguistic theories of politeness; one of their feature sets is a list of informal words derived from Wiktiontary, though this is not as helpful as punctuation and case features (possibly due to coverage). Other kinds of sociolinguistic text classification with involve stylistic variables include classification of gender (Garera and Yarowsky, 2009) and age (Rosenthal and McKeown, 2011).

Both the work of Michos et al. (1999) and that of Argamon et al. (2007) deserve special attention; they are similar in the sense that they both offer an extra level of interpretation between surface features and the classification of the text. Working in Greek, Michos et al. posit five discrete functional styles: *public affairs*, *scientific*, *journalistic*, *everyday*, and *literary*, each of which is characterized by a particular configuration of style features, i.e. *formality*, *elegance*, *syntactic complexity*, and *verbal complexity*, which have verbal and structural identifiers. For instance, public affairs and scientific texts have high *formality*, the verbal identifiers for which are formal words and lack of abbreviations and the structural identifiers for which are long sentences, fewer sentences per paragraphs, low verb/noun ratio, etc. Argamon et al. adopt attributes directly from Systemic Functional Grammar (Halliday, 1994), with the lexical features that trigger these attributes derived from the work of linguists in the SFG community (and expanded using WordNet and other lexical resources). They focus on three main *systems*: *Cohesion* (e.g. conjunctions), *Assessment* (e.g. modality), and *Appraisal* (e.g. positive or negative

attitude). Feature vectors for texts (or collections of texts) are constructed using the relative frequency of child options (e.g. *additive* conjunctions in general, rather than an instantiation such as *and*) to parent options (e.g. *extension*, or the grandparent *conjunction*). They show that such a feature vector can be used effectively in a number of text classification tasks: author identification, nationality identification, gender identification, personality typing, sentiment analysis, and identification of writing from different scientific fields. Though the usefulness of individual features varies depending on the task (and some features are even harmful in certain cases), the general conclusion is that this set of stylistic features clearly provides a useful 'toolbox' for various applications. This work highlights the potential of lexical resources for stylistic text classification.

Style is also relevant to readability classification. Si and Callan (2001) use a unigram Bayesian classifier to identify the reading level of scientific web pages in the K-8 range. Though the dataset is small, they are able to get markedly better performance than is possible with several of the most common metrics. Collins-Thompson and Callan (2005) expand this work to a wider grade range (K-13) and significantly larger dataset, applying more sophisticated statistical techniques. They do not outperform all of the simpler metrics on diagnostic reading level documents, but on web documents the smoothed unigram model is clearly superior. Tanaka-Ishii et al. (2010) address one obvious problem with the supervised approach: the need for texts annotated by all possible grade levels. They show how a general readability comparator, useful for any level of granularity, can be built from annotations for just two reading levels. Other work in readability has focused on increasing the feature space beyond unigrams; for instance Petersen and Ostendorf (2009) test SVM classifiers with a mixed bag of features, including basic text statistics (e.g. average sentence length), readability metrics, out-of-vocabulary scores, various parse features that indicate syntactic complexity, and perplexity features. Vajjala and Meurers (2012) offer a state-of-the-art model based on a wide variety of

syntactic and lexical features derived from Second Language Acquisition research.

4.1.2 Identifying stylistic inconsistency

The task of identifying stylistic inconsistency has only received a moderate amount of attention. Glover and Hirst (1996) ran a small experiment to collect two-part summaries, and then compared the first part of an essay written by one author with the second part written by another author. Follow-up work by Baljko and Hirst (1999), which had subjects sort various texts by style, showed that humans can consistently judge text style, but that human-judged style is not as closely tied to authorship as originally predicted. In more recent work, Graham et al. (2005) built artificial examples of style shift by concatenating consecutive Usenet postings by different authors, with all paragraph boundaries as potential style boundaries. Feature sets for their neural network classifiers included standard textual features, frequencies of function words, punctuation and parts of speech, lexical entropy, and vocabulary richness; however, only the frequency features proved to be useful, mostly because the other features were simply not effective for such small texts (i.e. two adjacent paragraphs). Nonetheless, the performance was well above baseline.

Guthrie (2008) presents some general methods for identifying anomalous segments within a larger text (or texts), and tests the effectiveness of his methods with a number of possible text variations: differing authorship, factual vs. opinion writing, news vs. subversive writing, and normal news vs. machine translated (Chinese) news. For testing, anomalous segments of varying length are inserted into other texts. For features, Guthrie uses simple textual features (e.g. word and sentence length), readability measures, obscure-vocabulary features, POS features (including a trigram diversity feature), frequency rankings of function words (which were not found to be useful), and context analysis features from the General Inquirer dictionary. The results suggest that statistics-based stylistic inconsistency detection is always difficult for small

spans of text where there is not enough information to be extracted.

Koppel et al. (2011) used a semi-supervised method to identify segments from two different books of the Bible artificially mixed into a single text. They first demonstrated that, in this context, preferred synonym use is a key stylistic feature that can serve as a high-precision bootstrap for building a supervised SVM classifier on more general features (common words); they then used this classifier to provide an initial prediction for each verse and smooth the results over adjacent segments. The method crucially relied on properties of the King James Version translation of the text in order to identify synonym preferences.

Stylistic inconsistency is a central theme within the task of intrinsic plagiarism detection (Stein et al., 2011), i.e. identifying plagiarism where the source is not available. The standard approach involves decomposition of the text, and then identification of outliers as defined by a stylistic feature model. Features for this task, mainly derived from work on authorship identification, include textual statistics, readability measures, POS and word features and, in particular, character trigrams. Stamatatatos (2009a) is notable for his use of a stylistic change function that does not assume a particular segmentation; his algorithm steps through the text and measures the difference between character trigram feature vectors at each point compared to the previous one to create a stylistic change function; maxima indicate potential instances of plagiarism. The approach of Kestemont (2011) decomposes the text into numerous overlapping spans, and then uses the similarity of character trigram feature vectors, refined using principal components analysis, to identify outliers.

4.1.3 Text generation

In the area of text generation, there is a general interest in making automatically generated texts sensitive to the contextual concerns of style. Early work by Hovy (1990) describes the pragmatics-modeling features of the text generator PAULINE, which makes stylistic choices,

when available, that satisfy its most pressing rhetorical goals. These goals are in turn are derived from a number of contextual features related to the conversational setting and participant characteristics, including their interpersonal goals. For instance, the desired formality of the text to be generated can be specified as a general contextual constraint, tone, or it can be derived from the relationship between the participants and the speaker's goals: if the (simulated) speaker is close to the hearer, or wishes to be, it will prefer colloquial language, unless the speaker dislikes the hearer(s), or wishes to anger them. If the interpersonal goal is to convince, then the situation is even more complex, the attitudes of both speaker and hearer towards the topic and their existing relationship will determine to what extent the bias is made explicit. Most of the stylistic goals have three possible settings, e.g. colloquial, normal, or highfalutin. Hard-coded strategies in six major areas (topic inclusion, topic organization, sentence organization, sentence constituent organization, and phrase/word choice) are followed to achieve these goals: for instance, formal texts are created by organizing topic by subordination, using adverbial groups, passive voice, complex verb tenses, nominalization, and using formal words while avoiding slang and contractions. This sort of approach depends explicitly on a mapping between stylistic goals and individual lexico-grammatical structures which must be identified prior to the use of the generation system.

The methodology of Paiva and Evans (2005) represents a modern, more statistical approach to style in text generation. They base their approach on the Multi-Dimensional (MD) analysis of Biber (1988): their text generator can be controlled to 'aim' for a particular style, as defined by a point in stylistic MD space (in a particular corpus; they test in the medical domain). This is accomplished by an offline-training mode where the generator first learns (by means of linear regression) how the particular choices it makes during the generation process will likely affect the ultimate stylistic output. In future runs, the generator can then greedily choose a path that is likely to lead to the desired stylistic effect, allowing for much quicker generation

than a more typical generation architecture, which would generate all possible outputs and then rank them. Comparing with other systems (Paiva and Evans, 2004), they note that this method offers stylistic flexibility not provided by other stochastic text generation systems, for instance HALogen (Langkilde-Geary, 2002), which tend to generate stylistically average (i.e. generic) texts. Although it is focused on simple textual features, such an approach could easily be expanded to include lexical stylistic features.

Other work that involves aspects of style relevant to text generation includes that of Inkpen and Hirst (2006); their lexical knowledge base, derived automatically from the near-synonym choice manual *Choose the Right Word* (Hayakawa, 1994), includes stylistic information such as formality, bias, and force. They integrate this information into a existing text generation system, HALogen, to improve individual lexical choice. The key difference as compared to PAULINE is that the stylistic tags are relative to other near-synonym choices, rather than absolutes, which offers an alternative way to inject stylistic information, provided you have an existing set of near-synonyms which can be contrasted for this purpose.

Many languages distinguish between informal and informal forms of address (e.g. *tu* or *vous*, T/V), but in other languages, like English, the distinction does not exist. Faruqui and Padó (2011) explore to what extent the T/V distinction can be recovered from the context in literary dialogue. Their prediction system relies heavily on the names of particular characters, which is unsatisfying even to the authors, but it highlights the kind of lexical choice task for which lexical stylistic information is particularly well suited. We will investigate a similar task in the next section.

4.1.4 Writing assistance

There is a huge body of literature on writing assistance but, except insofar as collocation and preposition errors can be viewed as stylistic rather than syntactic problems, style has received

relatively little attention in learner error correction. A notable early exception to this trend is the STASEL system of Payette and Hirst (1992), which performs a stylistic analysis of sentences with the goal of providing feedback to language learners. The syntactic style analyzer uses a full parse to identify problems with diction (too informal or clichéd writing), wordiness (redundancy forms and meaningless structures), structure (excessive noun modification, passive voice, ambiguity), and warnings for other questionable cases such as double negation. The syntactic and lexical frames that trigger feedback must be programmed manually; the need for lexical resources is clear. The other module in STASEL is a goal-directed style analyzer, providing information on goals like clarity and conciseness using the stylistic parser of DiMarco and Hirst (1993).

Most modern error-correction systems rely on some kind of error annotation. The only recent work on stylistic error annotation (of which I am aware) is that of Buscail and Saint-Dizier (2009). They present a taxonomy for stylistic errors in English, including errors of register (incorrect formality), deixis (failure to use relative time), coordination and reference (unresolvable ambiguity), tense (inconsistent tense), and sentence structure (lack of variation, too complex). After annotating a corpus according to their schema, they derive generalized correction rules, and show how they would present feedback to the user by means of an argumentation model that presents both the pros and cons for any particular change. At least one other recent FLL error annotation schema (Ramos et al., 2010) includes errors of register, though the primary focus of that work is collocation errors generally. Knowing which individual words and expressions belong to which registers can obviously provide a starting point for automatic identification of such errors.

Some error correction research leverages the major role that lexical L1 transfer plays in the writing errors of FLLs. The *Scripsi* system of Catt and Hirst (1990) specifically encodes information about French and Chinese that allows the system to identify particular errors in

the English writing of native speakers of those languages. Though mostly limited to grammar errors, the rules also encode information about idiomatic language that doesn't directly translate, e.g. *to have hunger* would be recognized as the direct translation of the French *avoir faim*, and then replaced with the colloquial English, *to be hungry*. Chang et al. (2008b) represents a modern, data-driven extension of this idea: in that work, native English data is used to identify the correct forms of various verb/noun combinations, but bilingual resources supply information about potential errors. For instance, the fact that the Chinese word *chi* is sometimes translated as *eat* and sometimes translated as *take* (as in *take medicine*) is first derived using a word-aligned Chinese-English corpus, and then when an anti-collocation *eat medicine* is found in Chinese learner text, alternative formulations of *eat* and *medicine* are generated, and the commonly appearing *take medicine* is suggested as an improvement; Dahlmeier and Ng (2011) present a generalized version of this idea that uses information from parallel corpora rather than bilingual lexicons. With respect to precision, this seems to be a better approach than looking for better collocation alternatives using lists of English synonyms (Futagi et al., 2008). From a lexical perspective, we might consider that certain terms could be identified as stylistically foreign, and, indeed, indicative of transfer from a particular language. We will return to this in Chapter 5.

A different but related kind of tool helps those writing for a language-impaired audience, with the goal of improving readability. For instance, Max (2006) presents an interactive system for text simplification: when a phrase matches a certain rule, the writer is immediately presented with one or more simpler alternatives (based on readability measures). The *Automated Text Adaption Tool* of Burstein et al (2007) provides relevant marginal notes and offers alternative vocabulary that might be easier, either because it is higher frequency or because it has a cognate in the native language of the students. Generally speaking, a stylistic lexicon should provide information about which words represent the stylistic fringes of the language,

and thus should be avoided; simple frequency, though a reasonable baseline, may be misleading in some cases, depending on the corpus that is being used as a reference.

The basic goal of an Automated Essay Scoring (AES) system is to mimic a human marker, analyzing a student essay and providing a rating that reflects its overall quality. Traditionally, the AES-graded essays were those produced as part of a standardized test, e.g. the GMAT (Burstein, 2003), though more recently AES has also been deployed in a general classroom setting (Warshauer and Ware, 2006; Scharber et al., 2008), providing feedback. The most common method for assessing the validity of AES is correlation with human graders, and from this perspective AES has been a resounding success: all of the major systems tend to agree with humans more than humans agree with each other (Keith, 2003). However, there are significant concerns that an automatic grading system must be held to a higher standard, namely *construct* validity (Chung and Baker, 2003): it is not enough that the scorer get the correct answer, it must do it for the right reasons. For instance, the scorer should not substitute an easily identified surface feature or proxy (e.g. text length), for a more subtle intrinsic variable (depth of analysis) just because the two are strongly correlated. This perceived failing, which allows for scorer to be fooled or 'spoofed' into giving an incorrect score, has led many to strongly criticize or outright reject AES (Cheville, 2004; Ericsson and Haswell, 2006). Using deeper lexical knowledge to replace the simple textual statistics may address some of these criticisms.

A general feature of the five AES systems discussed by Shermis and Burstein (2003) is some classification of the basic (surface) features into general categories such as mechanics, organization, content, and style; several systems allow for independent grading of these aspects (or *traits*), in addition to providing a *holistic* score. The most well-known AES system is probably *e-rater*, developed by the Educational Testing Services (Burstein et al., 1998; Burstein, 2003; Chodorow and Burstein, 2004; Attali and Burstein, 2006; Quinlan et al., 2009) In the most recent version of *e-rater* (V2.0) (Attali and Burstein, 2006) there are eight evaluated traits:

grammar, usage, mechanics, style, organization, development, lexical complexity, and use of prompt-specific vocabulary. For our purposes here we will focus on the style-relevant elements of the system, using the glossary of features from Quinlan et al. (2009). Most of the features included under the style heading are fairly simplistic: none require a parse, and the only one involving a lexicon is the use of 'inappropriate' words (e.g. profanity). Sentence construction is included in a very simple way: there is a feature for too many (more than 4) short sentences (less than 7 words), a feature for too many (more than 4) long sentences (more than 55 words), and a feature for too many sentences that begin with *and*. After an analysis suggesting that the effectiveness of the *short sentence* feature might be related to detection of fragments (under the grammar module), Quinlan et al. (2009) suggest that the feature should be removed. Another somewhat problematic feature is the passive voice, which only detects *by*-passives, and which was originally intended as a negative feature but in fact is correlated with a high style score (Quinlan et al., 2009).

There are other style-relevant *e-rater* features that are not included directly under the style heading. The lexical complexity trait has two features, sophistication of word choice and word length; the former is a metric based on word frequency. It is likely that some formality variation is being captured here. Similarly, the usage feature 'nonstandard verb or word form', which includes words associated with oral language, is also indicative of formality or a particular (non-standard) dialect. More generally, there is a serious question whether different 'traits' can be measured independently from one another using the standard AES approach (Lee et al., 2010), especially given that human judges, the judgments of whom are the basis of the system, tend to cluster their trait scores: for a given essay, agreement among various trait scores for a single judge tends to be higher than a single trait score given by different judges (Ponisciak and Johnson, 2003). The trend in *e-rater* research towards independent construct validity (Attali, 2007; Attali and Powers, 2007; Quinlan et al., 2009) has shown some successes in the areas of

mechanics and organization, but other syntactic and stylistic features, such as sentence variety, are missing (Lee et al., 2010).

4.2 Genre differentiation: Multidimensional Analysis vs. Latent Semantic Analysis

4.2.1 Introduction

As discussed in Section 2.2, multidimensional analysis (MD) is an approach popularized by Biber (1988) which uses dimensionality reduction on a mixed-genre to identify key, human interpretable dimensions of register.[1] Latent Semantic Analysis, as discussed in Section 3.2, is generally used for a very different purpose, modeling topic, and there are no claims generally made about the interpretability of individual dimensions of the model. However, it is not obvious that LSA cannot also be used to build a *register space*. Register space is the term I will use in this section refer to a vector space representation of register variation, consisting of a small number of real valued dimensions. Register spaces, as I define them, are usually derived in a bottom-up fashion from corpora that contain a variety of genres, such as the Brown and BNC. Each text is assigned a point in the space, with a good register space having different genres in different regions of the space. Another desirable quality of a register space is that the dimensions correspond to some human-identifiable characteristics of these texts. It is these two aspects that we will use to evaluate MD and LSA as methods for creating a register space.

There are two important aspects where MD and LSA differ. The first is feature set. The original Biber study used a small (56) set of features which were specifically selected to distinguish register. LSA, on the other hand, uses all of the words in the vocabulary as features.

[1] This work was originally presented as "Multidimensional analysis vs. latent semantic analysis for constructing a register space: Are hand-coded features needed, or is bag-of-words enough?" by Julian Brooke and Graeme Hirst, at the International Conference on Genre- and Register-related Text and Discourse Features in Multilingual Corpora (Brooke and Hirst, 2013b).

The second key difference is the dimensionality reduction technique. MD uses factor analysis, whereas the method underlying LSA is principal components analysis; they are related, but distinct.

4.2.2 Method

To replace the approach of multidimensional analysis, we implemented most of the features contained in the original Biber (1988) study, including textual statistics such as type-token ratio, POS-based features like various verb tenses, and also word classes as originally defined in Quirk et al (1985). There was a small subset of features that were ignored because they were difficult to derive accurately without a much more complex feature extraction system and/or human intervention: an example of this is omission of the *that* complementizer.

The standard approach to representing the features numerically in the multidimensional framework is to take their normalized frequency, that is frequency per 1000 tokens of text. With LSA, raw frequency is sometimes used; however, our work in lexicon acquisition (discussed in the previous Chapter) suggests that this is a poor choice when using open-class word features in the context of deriving style-relevant distinctions. Instead, for our BOW features we use a binary representation, that is a 0 or 1 depending on whether the word appeared in the text; we do not care how often it appeared.

Here, we consider two options for dimensionality reduction. As discussed earlier in Section 3.2, PCA, via SVD decomposition, provides a compression of a high-dimension space into a smaller one which is provably optimal in terms of retaining as much as the original variation as is possible within a given number of dimensions. It is a very popular technique in part because there are extremely fast algorithms for calculating its solution, which now allow for reduction of matrices with millions of rows and columns in a manageable time.[2] Factor analysis is a

[2]Here, as elsewhere, we used the divisi library for SVD, http://csc.media.mit.edu/divisi

related technique; it is distinct, though, because it does not try to represent the entire space, only common variation which are conceptualized as factors.[3] Some people have argued that this makes factor analysis more interpretable than PCA (Fabrigar et al., 1999). However, to our knowledge (and despite the similarity of the methods) the kinds of optimizations that are available for PCA are not directly applicable to factor analysis, and so factor analysis is usually applied only with relatively small data or feature sets, and is not used much at all in the NLP community at large.

Our experiments will include both a qualitative and quantitative component. We will be primarily offering an analysis of the specific distribution of spaces created by our techniques, but we also want a simple numerical metric that reflects whether our register space successfully distinguishes among genres. *Register differentiation* does this by comparing the distances between the averages of pairs of genres for each dimension, with the distance normalized by the standard deviation. Given a set of known genres G whose mean and standard deviation relative to a set of dimensions D is defined, we further define register differentiation (RD) as follows:

$$RD(G,D) = \frac{\sum_{d \in D} \sum_{g_1 \in G} \sum_{g_2 \in G} |\frac{\overline{g_{1d}} - \overline{g_{2d}}}{\sigma_{g_{1d}}}|}{|D||G|^2}$$

Thus, register differentiation is high when there is low within genre variation, and high cross-genre variation.

4.2.3 Dimensionality reduction experiment

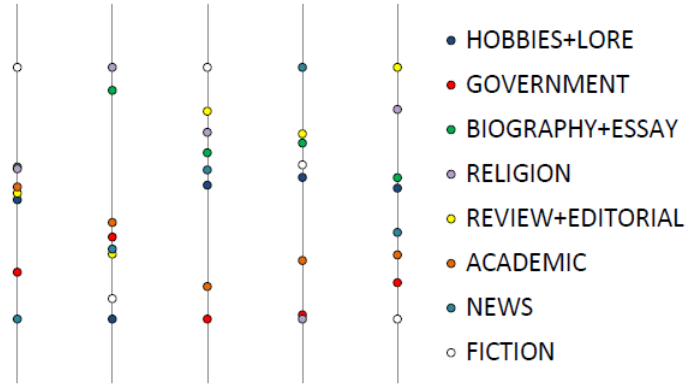
Our first experiment looks at whether the kind of dimensionality reduction chosen has a major effect on the character of the register space created. For this experiment, we use the multi-dimensional feature set in the Brown corpus, and test using factor analysis or principal components analysis. In this work, we will use 5 dimensions in the final space, which is a fairly

[3] For factor analysis we used the mdp toolkit, http://mdp-toolkit.sourceforge.net/

Table 4.2: Register differentiation for dimensionality reduced register spaces in the Brown Corpus. MD features are used.

Type of Dimensionality Reduction	Register Differentiation
Factor Analysis	0.81
Principal Components Analysis	0.92

Figure 4.1: Register dimensions for factor analysis, MD features, Brown Corpus.

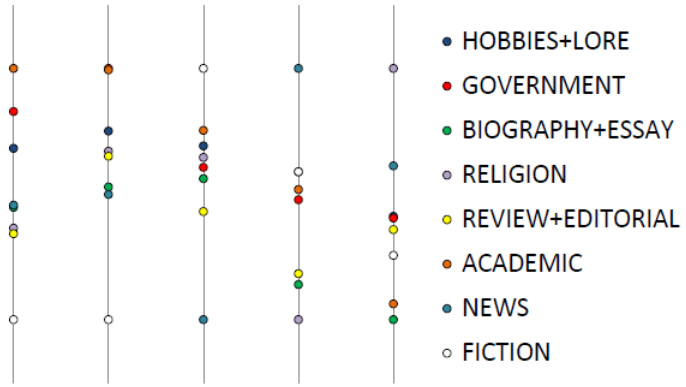


common number in MD work and is small enough that it is possible to offer a tentative qualitative analysis.

The quantitative results are in Table 4.2. We found that PCA was a bit better than factor analysis with respect to our register differentiation metric, though the difference is fairly small.

In Figure 4.1 we see the 5 dimensions for factor analysis; the vertical positions of the dots correspond to the average values of selected genres in the Brown. All the dimensions have been scaled so we can focus on the variation within dimensions, rather than the individual numerical values. We will refer to dimension 1 as the one on the far left. One obvious result is that, for 4 of the 5 dimensions, fiction is at or near one extreme. This is not a surprise, since in the entirely written Brown corpus fiction is likely to be the most "spoken", and therefore interactional rather than informational; we would also expect it to be the most narrative, and perhaps the least abstract; observe the opposition to religious, government, and academic documents. Dimensions 4 and 5, defy a clear interpretation.

Figure 4.2: Register dimensions for PCA, MD features, Brown Corpus.



For PCA results in Figure 4.2, the story is similar: again, fiction stands out in these low dimensions, with even clearer opposition to the same genres as before: in dimension 2, though, news is the closest to fiction, suggesting a narrative dimension, but in dimension 3 they are fully opposed, again indicating an informational/interactive distinction, perhaps. Dimension 4 has news at one end strongly opposed to reviews, essays and religion; a subjective or persuasive aspect of the texts might be the key distinction. Dimension 5, though, makes little sense based purely on the distribution of genres.

So, in this first experiment, we found that PCA offered a slightly better register differentiation with respect to our metric, and the quantitative analysis for PCA was not much different than factor analysis, with lots of consistency with the kinds of dimensions identified in Biber's work.

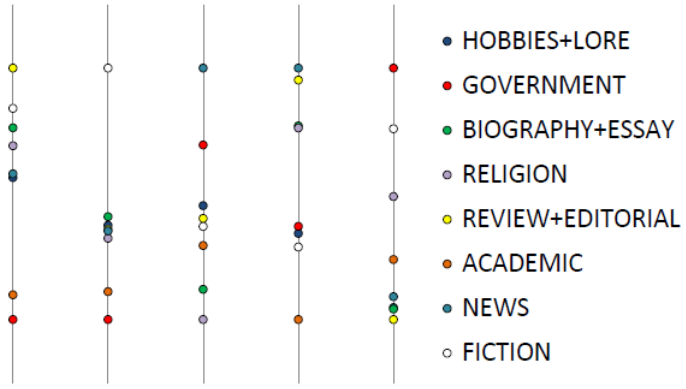
4.2.4 Feature set experiments

For our second experiment, we use only PCA dimensionality reduction, and focus on our key interest: the use of traditional MD features versus a bag-of-words approach. We carried out this experiment in two corpora, both the Brown as well as the British National Corpus, which has a much wider range of genres, including a significant number of spoken genres.

Table 4.3: Register differentiation for PCA Dimensionality reduction

Feature set	Register Differentiation	
	Brown	BNC
MD features	0.92	1.81
Bag-of-words	1.29	2.11

Figure 4.3: Register dimensions for LSA (BOW), Brown corpus



The qualitative results are in Table 4.3. For both corpora, the binary bag-of-words approach offers much better genre differentiation than MD features. The differences here are more striking than the differences between dimensionality reduction techniques.

For the Brown corpus, the qualitative analysis for the MD features with PCA is in Figure 4.2. For LSA in the Brown (Figure 4.3), Dimension 2 is the most clear cut, with an extreme opposition between fiction and more abstract, informational genres like government and academic documents; on close inspection, we see that Dimension 1 is similar though the position of review/editorial and essays along with fiction at one end may indicate a subjective aspect. Dimensions 3 and 4, with an opposition between news and religious, essays, and academic documents, suggests a distinction between plain, just-the-facts language, and fancier language, either literary or scientific/jargony. The 5th dimension is harder to categorize though its endpoints are the same as the first. All in all, LSA seems to offer more variety of stylistic aspects, but not obviously less interpretability.

Figure 4.4: Register dimensions for MD features, BNC

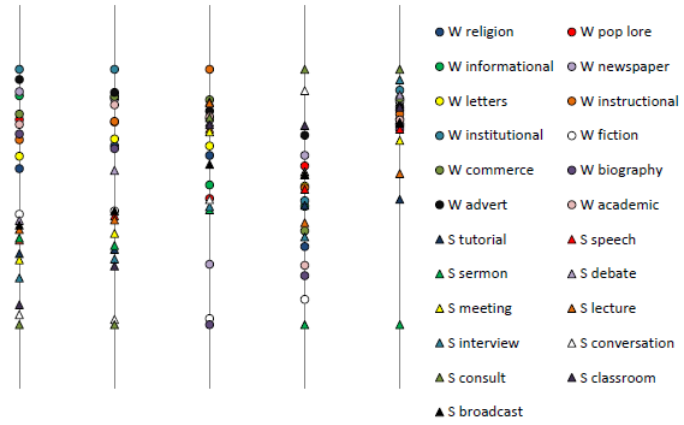
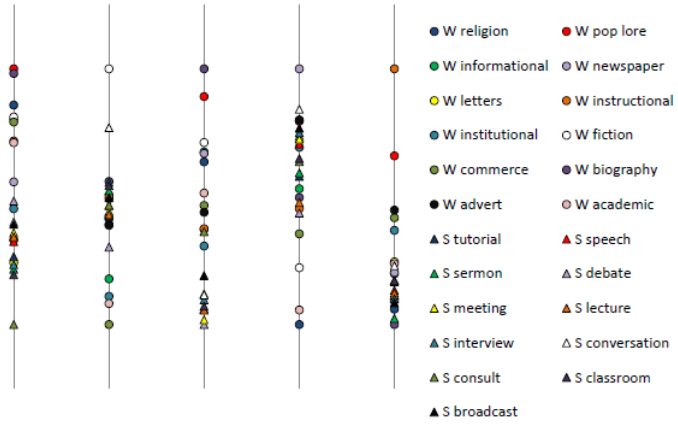


Figure 4.5: Register dimensions for LSA (BOW), BNC



Turning now to the BNC, we see in Figure 4.4 that with MD features both of the first two dimensions do a reasonably good job of distinguishing spoken genres (triangles) from written genes (circles). As expected, fiction is the borderline case. In the 3rd dimension, there is a contrast between narrative genres and all the others. Sermons and consultations form the extreme of both the 4th and 5th dimensions, perhaps indicating an interactional aspect.

Using a bag of words feature set (Figure 4.5), we also see two dimensions (the 1st and 3rd) which fairly consistently divide written and spoken genres, though it is somewhat less categorical. Dimension 2 seems to be distinguishing fairly concrete genres (fiction, conversation) from much more abstract, esoteric ones (commerce, academic, and government writings,

as well as debates). Dimension 4 also seems to distinguish everyday, general consumption genres, including news, from fancier language (academic, religious writing).

Our second experiment indicates that from a purely quantitative perspective, BOW distinguishes genres better. Qualitatively, all of the conditions offer some interpretable, reasonable dimensions, and also others that appear repetitive or are entirely uninterpretable. In both cases, we saw distinctions between writing and speaking, though the MD features perhaps had the upper hand in that regard (and indeed, that was what they were originally designed to distinguish). We conclude, tentatively, that a BOW approach can be used to build a "good" register space, distinct but ultimately similar to an MD approach. But we do have a major concern: using so many features in the LSA approach, could we be distinguishing these genres in a trivial way that has more to do with the specific choices made by the creators of these corpora, rather than getting at real distinctions between genres? In short, are we overfitting?

4.2.5 Cross-space experiments

Experiment 3 tests the robustness of our register spaces created using LSA. Within the mathematical representation provided by LSA, it is easy to add new texts into a register space created using another set of texts; if we returning to the original SVD breakdown into U, Σ, and V^T (see Section 3.2) multiplying a new text vector (in the original feature representation) by V^T will yield a vector in the same format as the rows of U (though the optimality of the representation is not preserved). Here, we compare Brown texts within a "Brown space" to these same texts transferred into the BNC space from experiment 2.

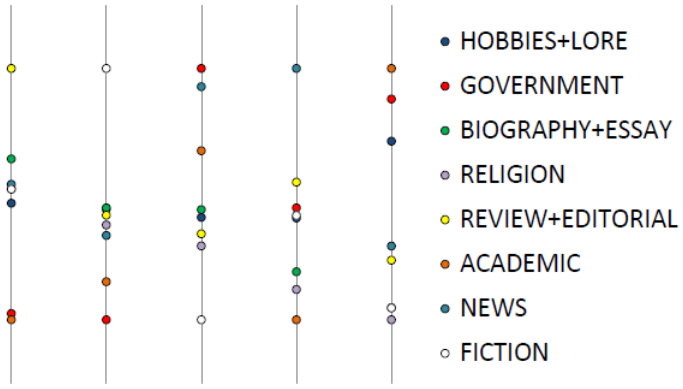
Table 4.4 indicates that the register differentiation between the two spaces is roughly equivalent, though the Brown is slightly better in its own register space.

Figure 4.6 shows the LSA-derived dimensions for the Brown texts in BNC space. Again, fiction stands out in dimensions 2 and 3 as compared to more abstract, less interactional and

Table 4.4: Register differentiation for LSA, Brown texts

Register Space Corpus	Register Differentiation
Brown	1.29
BNC	1.23

Figure 4.6: Register dimensions for LSA, Brown corpus, BNC register space



narrative genres. With no spoken texts included, Dimension 1 again has a subjective bent, with reviews and essays at one end and more objective government and academic documents at the other, and Dimension 4 continues to represent everyday vs fancier language, technical or literary, and dimension 5 distinguishes more literary-oriented texts. All in all, this cross-corpus method seems to result in more rather than less interpretable dimensions, with less repetition.

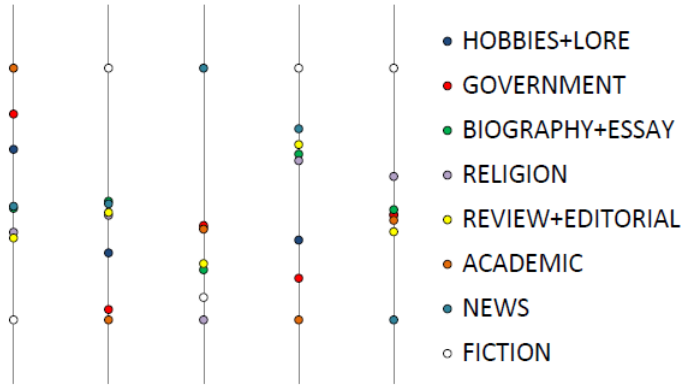
Our final experiment also involves a change in space, though here we consider the space of features, rather than the space of texts. If we define features in terms of texts (as we did during our induction experiments), we can transfer features into a new feature space, just as we did with texts in the previous experiment. Using the Brown corpus, we build a feature register space for each of our two feature sets. We can turn a feature register space into a text register space by simply summing across all the features in a text to get a vector representation for the whole text. We compare the MD features within their original space with the MD features transferred into the LSA space.

The quantitative results in Table 4.5 offer two interesting conclusions. Interestingly, the

Table 4.5: Register differentiation for feature types, Brown texts

Register Space Creation Method	Register Differentiation
Multidimensional Analysis	0.92
Latent Semantic Analysis	1.04

Figure 4.7: Register dimensions for MD features in LSA space, Brown corpus



MD features do slightly better in the LSA space than they do in their original space. This would seem to confirm that the LSA space is a "better" space, it actually provides a boost to the MD features in terms of their ability to distinguish genres. One other very nice result here is that the register differentiation in the purely MD space using our roundabout feature space method is exactly identical to the original differentiation in the text register space. This suggests (though we have not mathematically proven this) that text register spaces and feature register spaces are fully interchangeable; we can focus on building the latter, and then use those register spaces to build the former on the fly, with similar results to a text-focused approach.

Qualitatively, the results for the MD features in LSA space are very similar to the original ones for the MD-based space, though the distinction between fiction and other genres is actually larger than we saw with either MD features or the LSA space individually.

In these two experiments, we have tested transferring texts and features into "foreign" register spaces, i.e. register spaces which did not originally include these texts or features. This appears to preserve most of the qualities we would expect, suggesting the spaces we are build-

ing are fairly robust. More generally, we conclude that LSA, both in terms of the PCA dimensionality reduction used and its bag-of-words feature set, seems to be a good way to approach register differentiation, resulting in dimensions which both distinguish genres quantitatively and offer human interpretable dimensions.

4.3 Lexical sociolinguistics

In this section, I present the results of a small experiment testing whether the formality score metric, derived in Section 3.4, is useful for distinguishing sociolinguistic factors in the Toronto corpus (Tagliamonte, 2006), which was used for the sociolinguistic variable extraction method in Section 3.8.[4] The methodology for the main experiment is quite simple: for each text (interview) in the Toronto corpus, we assigned a formality score based on the average formality score of all the words in the text. Then we divided the texts into groups based on three sociolinguistic factors: age, type of work (blue collar or white collar), and gender, and compared average formality for various factors. We also ran t-tests to see if the differences were statistically significant, and looked at individual words which contributed most overall to the differences.

Figure 4.8 contains the results for age, broken down into age bands. Note that formality is consistently negative in this corpus, which is expected given that these are spoken texts. However, there are clear differences by age, with a pattern of increasing formality with older speakers. The differences between those older than 30 and less than 30 are statistically significant ($p < 0.001$), as are those between children (18 and younger) and young adults (between 19 and 30) ($p < 0.01$). Some of the words that had the most influence on formality here were low formality words used by the young such as *like*, *yeah*, and *stuff*.

Figure 4.9 contains the results for type of work. Again, the differences are fairly stark,

[4]The work in this section was presented as "Facets of formality: A dimension of register in a sociolinguistic corpus" by Julian Brooke and Graeme Hirst, at the Georgetown University Round Table on Languages and Linguistics 2012 (Brooke and Hirst, 2012a).

Figure 4.8: Average formality by age

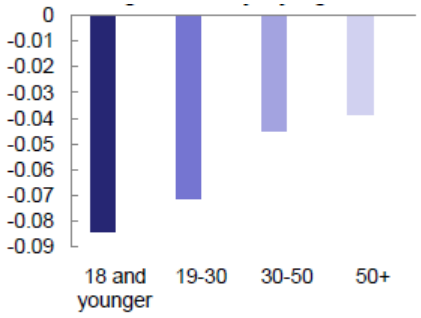
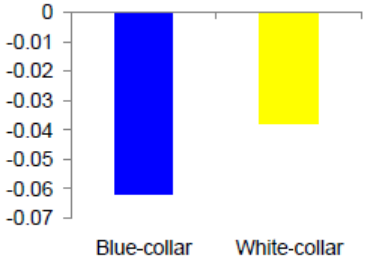


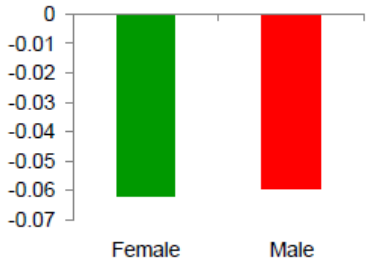
Figure 4.9: Average formality by economic class



with blue-collar workers tending to be much more informal than white-collar workers. This difference is significant ($p < 0.001$). Low-formality items preferred by blue-collar workers include *gotta*, *stuff*, and *guy*, while white-collar workers used the intensifier *very*, which as we've already discussed is becoming more formal. By contrast, we did not see any significant differences by the gender of speaker, which is shown in Figure 4.10. Note that we did find some differences in the particular informal words that men and women used; for example, men preferred *gotta* while women preferred *oh-my-god* (which is consistently hyphenated as one word in the Toronto corpus).

It is important to note that our conception of formality does not depend directly on social factors like age: our notion of 'appropriateness' from Section 3.4 was grounded in differences in medium and genre, and no such differences exist in the Toronto corpus. Nevertheless, the results presented above are not terribly surprising in that they seem to be confirming basic

Figure 4.10: Average formality by gender

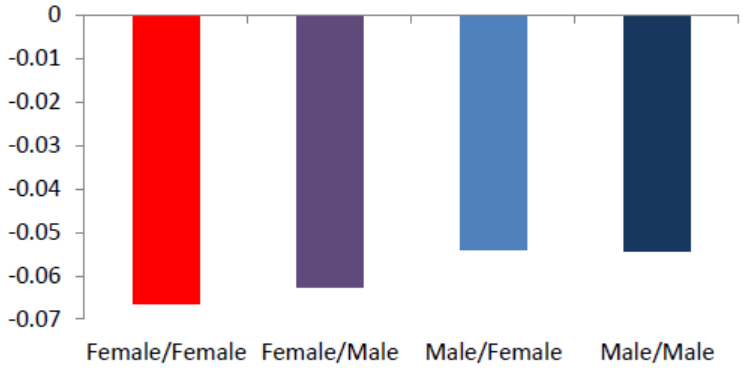


stereotypes about how social factors and styles are related. Our last experiment in this section, though, involves a less obvious result. To this point, we have only considered the idea that the social demographics of the speaker (the interviewee) has a direct influence on the level of formality; we have not considered the social demographics of the listener (the interviewer). As it happens, the interviewers for the Toronto corpus were all undergraduate students in their early 20s (Tagliamonte, personal communication), so we cannot separate out any effect for interviewer age and social class: there is, however, some variation with respect to interviewer gender, and this information is included in the corpus.

In Figure 4.11, we break down the gender results in Figure 4.10 to show the average formality score both by interviewer and interviewee gender. Although we did not find a difference based on interviewee gender, we do see one based on interviewer gender: both men and women spoke more informally to women.[5] This result, which intuitively could be attributed to real social differences between genders (we will not speculate further here to the exact cause), is not as strong as some of those based on interviewer social factors, but it is statistically significant ($p < 0.05$). Another possibility, that interviewees would talk more informally to those of the same gender (i.e. less social distance), regardless of specific gender, was not borne out; though there is a small tendency in that direction, it is not statistically significant and is entirely at-

[5]We were concerned that this might be an artifact of the way the corpus was collected—perhaps there is a trivial relationship between interviewer gender and other social factors of the interviewee which we already know have influence—so we confirmed that this effect seems to hold among social factor subgroupings of interviewees, though we do not present those results here.

Figure 4.11: Average formality by gender pairing (interviewer/interviewee)



tributable to increased informality in just the female-female case (the most informal). The idea that the gender of the partner has a major effect on the lexical choices of the speaker is not a new one (Garera and Yarowsky, 2009), but making this connection, in combination with the more straightforward results presented here, strongly suggests how our 'fuzzy' stylistic variables can be applied to a variety of more-concrete profiling tasks.

4.4 Word clipping prediction

4.4.1 Introduction

Clipping is a type of word formation where the beginning and/or the end of a longer word is omitted (Kreidler, 1979).[6] This phenomenon is attested in various languages; well-known examples in English include words such as *hippo* (*hippopotamus*) and *blog* (*weblog*). Clipping and related kinds of word formation have received attention in computational linguistics with respect to the task of identifying source words from abbreviated forms, which has been studied, for instance, in the biomedical and text messaging domains (Okazaki and Ananiadou, 2006; Cook and Stevenson, 2009).

[6]The work in this section is based on "Clipping prediction with latent semantic analysis" by Julian Brooke, Tong Wang, and Graeme Hirst, published in the *Proceedings of the 5th International Joint Conference on Natural Language Processing* (Brooke et al., 2011).

Compared to many near-synonyms, clipped forms have the important property that the differences between full and abbreviated forms are almost entirely connotational or stylistic, closely tied to the formality of the discourse.[7] This fact allows us to pursue two distinct though related approaches to this task, comparing a supervised model of word choice (Wang and Hirst, 2010) with a mostly unsupervised system that leverages an automatically-built lexicon of formality (Brooke et al., 2010b). Our findings indicate that the lexicon-based method is highly competitive with the supervised, task-specific method. Both models approach the human performance evidenced in an independent crowdsourced annotation.

4.4.2 Methods

Both approaches that we are investigating make use of Latent Semantic Analysis (LSA) as a dimensionality-reduction technique (Landauer and Dumais, 1997). Our first method is the lexical choice model proposed by Wang and Hirst (2010). This approach performs SVD on a term–term co-occurrence matrix, which has been shown to outperform traditional LSA models that use term–document co-occurrence information. Specifically, a given word w is initially represented by a vector v of all its co-occurring words in a small collocation context (a 5-word window), i.e., $v = (v_1, \ldots, v_n)$, where n is the size of the vocabulary, and $v_i = 1$ if w co-occurs with the i-th word in lexicon, or $v_i = 0$ otherwise. The dimensionality of the original vector is then reduced by SVD.

A context, typically comprising a set of words within a small collocation context around the target word for prediction (though we test larger contexts here), is represented by a weighted centroid of the word vectors. Together with the candidate words for prediction, this context vector can then be used as a feature vector for supervised learning; we follow Wang and Hirst in using support vector machines (SVMs) as implemented in WEKA (Witten and Frank, 2005),

[7] Shortened forms might also be preferred in cases where space is at a premium, e.g. newspaper headlines or tweets.

training a separate classifier for each full/clipped word form pair. The prediction performance varies by k, which can be tested efficiently by simply truncating a single high-k vector to smaller dimensions. The optimal k value reported by Wang and Hirst testing on a standard set of seven near-synonyms was 415; they achieved an accuracy of 74.5%, an improvement over previous statistical approaches, e.g. Inkpen (2007).

The competing method involves building lexicons of formality, using the method discussed in Section 3.4, which is itself an adaptation of an approach used for sentiment-lexicon building (Turney and Littman, 2003). Though it relies on LSA, there are several key differences as compared to the context vector approach. First, the pre-LSA matrix is a binary word–document matrix, rather than word–word. For the LSA step, we showed that a very low k value (20) was an appropriate choice for identifying variation in formality. After dimensionality reduction, each word vector is compared, using cosine similarity, to words from two sets of seed terms, each representing prototypical formal and informal words, which provides a formality score for each word in the range of -1 to 1. See Section 3.4 for more about the derivation of the final formality score. Our evaluation suggests that, given a large-enough blog corpus, this method almost perfectly distinguishes words of extreme formality, and is able to identify the more formal of two near-synonyms over 80% of the time, better than a word-length baseline.

Given a lexicon of formality scores, the preferred form for a context is identified by averaging the formality scores of the words in the context and comparing the average score to a cutoff value. Here, the context is generally understood to be the entire text, though we also test smaller contexts. We take the cutoff to be midpoints of the average scores for the contexts of known instances; although technically supervised, we have found that in practice just a few instances is enough to find a stable, high-performing cutoff. Note that the cutoff is analogous to the decision hyperplane of an SVM. In our case, building a lexical resource corresponds to additional task-independent reduction in the dimensionality of the space, greatly simplifying

the decision.

4.4.3 Resources

Blog data is an ideal resource for this task, since it clearly contains a wide variety of language registers. For our exploration here, we used a collection of over 900,000 blogs (216 million tokens) originally collected from the web in May 2008. We segmented the texts, filtered out short documents (less than 100 words), and then split the corpus into two halves, training and testing. For each of the two methods described in the previous section, we derived the corresponding LSA-reduced vectors for all lower-case words using the collocation information contained within the training portion.[8] The testing portion was used only as a source for test contexts.

We independently collected a set of common full/clipped word pairs from web resources such as Wikipedia, limiting ourselves to phonologically-realized clippings. This excludes orthographic shortenings like *thx* or *ppl* which cannot be pronounced. We also removed pairs where one of the words was quite rare (fewer than 150 tokens in the entire corpus) or where, based on examples pulled from the corpus, there was a common confounding homonym–for instance the word *prob*, which is a common clipped form of both *problem* and *probably*. However, we did keep words like *doc*, where the *doctor* sense was much more common than the *document* sense. After this filtering, 38 full/clipped word pairs remained in our set. For each pair, we automatically extracted a sample of usage contexts from texts in the corpus where only one of the two forms appears. For each word form in each of our training and testing corpora, we manually removed duplicate and near-duplicate contexts, non-English and unintelligible contexts, and any remaining instances of homonymy until we had 50 acceptable usage exam-

[8] We used the same dataset for each method so that the difference in raw co-occurrence information available to each method was not a confounding factor. However, we also tested the lexicon method using the full formality lexicon from Section 3.4, built on the larger ICWSM blog corpus; the difference in performance was negligible.

ples for each word form in each sub-corpus (100 for each of the word pairs), a total of 3800 contexts for each of training and testing.

One gold standard is provided by the original choice of the writer, but another possible comparison is with reference to an independent human annotation, as has been done for other near-synonym word choice test sets (Inkpen, 2007). For our annotation, we used the crowdsourcing website Crowdflower (www.crowdflower.com), which is built on top of the well-known Amazon Mechanical Turk (www.mturk.com), which has been used, for instance, to create emotion lexicons (Mohammad and Turney, 2010). In general terms, these crowdsourcing platforms provide access to a pool of online workers who do small tasks (HITs) for a few cents each. Crowdflower, in particular, offers a worker-filtering feature where gold standard HITs (75 clear instances taken from the training data) are interspersed within the test HITs, and workers are removed from the task if they fail to answer a certain percentage correct (90%). For each word form, we randomly selected 20 of 50 test contexts to be judged, or 1520 altogether. For each case, the workers were presented with the word pair and three sentences of context (additional context was provided if less than 40 tokens), and asked to guess which word the writer used. To get more information and allow participants to express a tentative opinion, we gave the workers five options for a word pair A/B: "Probably A/B", "Definitely A/B", and "I'm not sure"; for our purposes here, however, we will not distinguish between "Probably" and "Definitely". We queried for five different judgments per test case in our test corpus, and took the majority judgment as the standard, or "I'm not sure" if there was no majority judgment.

4.4.4 Evaluation

First, we compare our crowdsourced annotation to our writer's choice gold standard, which provides a useful baseline for the difficulty of the task. The agreement is surprisingly low; even if "I'm not sure" responses are discounted, agreement with the writer's choice gold standard

Table 4.6: Clipping prediction results, all pairs

Vector classification	
Options	**Accuracy (%)**
Sentence context only (k=17)	62.9
3-sentence context (k=15)	64.6
Full document (FD) (k=16)	**67.9**
FD, single (generalized) classifier	66.8
FD, best k for each pair	65.9
Formality lexicon	
Options	**Accuracy (%)**
Sentence context only	65.2
3-sentence context	65.5
Full document (FD)	**66.7**
FD, single (generalized) cutoff	65.1

is just 71.7% for the remaining datapoints. For certain words (such as *professor*, *doctor*), workers avoided the non-standard clipped forms almost entirely, though there were other pairs, like *photo/photograph*, where the clipped form dominated. Expected frequency, rather than document context, is clearly playing a role here.

Our main evaluation consists of comparing the predictions of our two methods to the original choice of the writer, as seen in our corpus. Accuracy is calculated as the number of predictions that agree with this standard across all the (3800) contexts in our test set. We first calibrated each model using the training set, and then prompted for predictions with various amount of context.[9] The 3-sentence context includes the sentences where the word appeared, and the sentences on either side. Other options we investigated were, for the vector classification, the option of using a single classifier for all pairs, or using a different k-value for each pair, and, for the lexicon-based prediction, the option of using a single cutoff for all pairs. The best k were determined by 10-fold cross-validation on the training set. The results are given in Table 4.6. Since our test sets are balanced, the random guessing performance is 50%.

A chi-square test indicates the difference between the best performing result for each

[9]In all cases, other appearances of the word or an inflected form in the context were removed.

method is not statistically significant. We see that both methods show an improvement with the addition of context beyond the sentence where the word appears, with full document context providing the best results; the improvement with full document context is statistically significant for the vector classification model ($p < 0.001$). Overall, the two methods make similar choices, with the agreement of the predictions at 78.1% for the full document models. Another result that points to the similarity of the final models is that the best single k value is very close to the best k value for formality lexicon building. The generalized clipping models (of both kinds) do worse than the pair-specific models, but the drop is fairly modest. An even more individualized vector classification model, in the form of individual k values for each pair, does not improve performance. If we instead take the worker judgements as a gold standard, the performance of our two models on that subset of the test data is worse than with a writer-based standard: 61.1% for the best lexicon-based model, and 63.6% for the best vector classification model.

Finally, we look at individual full/clipped word pairs. Table 4.7 contains the results for a sample of these pairs, using the best models from Table 4.6. Some word pairs (e.g. *mic/microphone*) were very difficult for the models, while others can usually be distinguished. The main difference between the two models is that there are certain pairs (e.g. *plane/airplane*) where the vector classification works much better, perhaps indicating that formality is not the most relevant kind of variation for these pairs.

4.4.5 Discussion

Our initial hypothesis was that the formality of the discourse plays a key role in determining whether a clipped form of a word will be used in place of a full form, and thus a lexicon of formality could be a useful tool for this kind of word choice. Our results mostly bear this out: although the vector classification model has a slight advantage, the lexicon-based method,

Table 4.7: Clipping prediction results, by pair

	Accuracy (%)	
Clipped pair	**VC model**	**FL model**
prof/professor	68	74
tourney/tournament	64	55
plane/airplane	61	42
doc/doctor	81	78
stats/statistics	74	75
meds/medication	82	82
fridge/refrigerator	65	63
app/application	66	62
mic/microphone	54	59
fam/family	84	85

which has the advantage of compactness, interpretability, and portability, does reasonably well. Tellingly, the best vector-based model is very similar to the lexicon in terms of its parameters, including a preference for the use of the entire document as context window and low LSA k, rather than the local context and high LSA k that was preferred for a previous near-synonym choice task (Wang and Hirst, 2010). In comparison to that task, clipping prediction is clearly more difficult, a fact that is confirmed by the results of our crowdsourced annotation.

The fact that the models do better on certain individual word pairs and more poorly on others indicates that the degree of formality difference between clipped and full forms is probably quite variable, and in some cases may be barely noticeable. Under those circumstances, the advantages of a vector classification model, which might base the classification on other kinds of relevant context (e.g. topic), are clear. We conclude by noting that for a highly specialized problem such as word clipping prediction, a single lexical resource can, it appears, complete with a task-based supervised approach, but even here we see signs that a single resource might be insufficient to cover all cases. For wider, more complex tasks, any particular resource may address only a limited part of the task space, and therefore a good deal of work may be required before a lexicon-based method can reasonably compete with a more straightforward statistical

approach.

4.5 Stylistic segmentation in *The Waste Land*

4.5.1 Introduction

Most work in automated stylistic analysis operates at the level of a text, assuming that a text is stylistically homogeneous.[10] However, there are a number of instances where that assumption is unwarranted. One example is documents collaboratively created by multiple authors, in which contributors may, either inadvertently or deliberately (e.g. Wikipedia vandalism), create text which fails to form a stylistically coherent whole. Similarly, stylistic inconsistency might also arise when one of the 'contributors' is actually not one of the purported authors of the work at all — that is, in cases of plagiarism. More-deliberate forms of stylistic dissonance include satire, which may first follow and then flout the stylistic norms of a genre, and much narrative literature, in which the author may give the speech or thought patterns of a particular character their own style distinct from that of the narrator. In this subsection, we address this last source of heterogeneity in the context of the well-known poem *The Waste Land* by T.S. Eliot, which is often analyzed in terms of the distinct voices that appear throughout the text.

T.S. Eliot (1888–1965), recipient of the 1948 Nobel Prize for Literature, is among the most important twentieth-century writers in the English language. Though he worked in a variety of forms — he was a celebrated critic as well as a dramatist, receiving a Tony Award in 1950 — he is best remembered today for his poems, of which *The Waste Land* (1922) is among the most famous. The poem deals with themes of spiritual death and rebirth. It is notable for its disjunctive structure, its syncopated rhythms, its wide range of literary allusions, and its incorporation of numerous other languages. The poem is divided into five parts; in total it is

[10]The work presented in this section was adapted from "Unsupervised stylistic segmentation of poetry with change curves and extrinsic features" by Julian Brooke, Adam Hammond, and Graeme Hirst, published in the *Proceedings of the 1st Workshop on Computational Literature for Literature* (Brooke et al., 2012a).

433 lines long, and contains 3533 tokens, not including the headings.

A prominent debate among scholars of *The Waste Land* concerns whether a single speaker's voice predominates in the poem (Bedient, 1986), or whether the poem should be regarded instead as dramatic or operatic in structure, composed of about twelve different voices independent of a single speaker (Cooper, 1987). Eliot himself, in his notes to *The Waste Land*, supports the latter view by referring to "characters" and "personage[s]" in the poem.

One of the poem's most distinctive voices is that of the woman who speaks at the end of its second section:

> I can't help it, she said, pulling a long face,
>
> It's them pills I took, to bring it off, she said
>
> [158–159]

Her chatty tone and colloquial grammar and lexis distinguish her voice from many others in the poem, such as the formal and traditionally poetic voice of a narrator that recurs many times in the poem:

> Above the antique mantel was displayed
>
> As though a window gave upon the sylvan scene
>
> The change of Philomel
>
> [97–99]

While the stylistic contrasts between these and other voices are apparent to many readers, Eliot does not explicitly mark the transitions between them. The goal of the present work is to investigate whether computational stylistic analysis can identify the transition between one voice and the next.

Our unsupervised approach, informed by research in topic segmentation (Hearst, 1994) and intrinsic plagiarism detection (Stamatatos, 2009a), is based on deriving a curve representing

stylistic change, where the local maxima represent likely transition points. Notably, our curve represents an amalgamation of different stylistic metrics, including those that incorporate external (extrinsic) knowledge, e.g. vector representations based on larger corpus co-occurrence, which we show to be extremely useful. For development and initial testing we follow other work on stylistic inconsistency by using artificial (mixed) poems, but the our main evaluation is on *The Waste Land* itself. We believe that even when our segmentation disagrees with expert human judgment, it has the potential to inform future study of this literary work.

4.5.2 Related work

Poetry has been the subject of extensive computational analysis since the early days of literary and linguistic computing (e.g., Beatie 1967). Most of the research concerned either authorship attribution or analysis of meter, rhyme, and phonetic properties of the texts, but some work has studied the style, structure, and content of poems with the aim of better understanding their qualities as literary texts. Among research that looks at variation with a single text, Simonton (1990) found quantitative changes in lexical diversity and semantic classes of imagery across the components of Shakespeare's sonnets, and demonstrated correlations between some of these measures and judgments of the "aesthetic success" of individual sonnets. Duggan (1973) developed statistical measures of formulaic style to determine whether the eleventh-century epic poem *Chanson de Roland* manifests primarily an oral or a written style. Also related to our work, although it concerned a novel rather than a poem, is that of McKenna and Antonia (2001), who used principal component analysis of lexical frequency to discriminate different voices (dialogue, interior monologue, and narrative) and different narrative styles in sections of *Ulysses* by James Joyce.

Topic segmentation is a similar problem that has been quite well-explored. A common thread in this work is the importance of lexical cohesion, though a large number of compet-

ing models based on this concept have been proposed. One popular unsupervised approach is to identify the points in the text where a metric of lexical coherence is at a (local) minimum (Hearst, 1994; Galley et al., 2003). Malioutov and Barzilay (2006) also used a lexical coherence metric, but applied a graphical model where segmentations are graph cuts chosen to maximize coherence of sentences within a segment, and minimize coherence among sentences in different segments. Another class of approaches is based on a generative model of text, for instance HMMs (Blei and Moreno, 2001) and Bayesian topic modeling (Utiyama and Isahara, 2001; Eisenstein and Barzilay, 2008); in such approaches, the goal is to choose segment breaks that maximize the probability of generating the text, under the assumption that each segment has a different language model.

4.5.3 Stylistic change curves

Many popular text segmentation methods depend crucially on a reliable textual unit (often a sentence) which can be reliably classified or compared to others. But, for our purposes here, a sentence is both too small a unit — our stylistic metrics will be more accurate over larger spans — and not small enough — we do not want to limit our breaks to sentence boundaries. Generative models, which use a bag-of-words assumption, have a very different problem: in their standard form, they can capture *only* lexical cohesion, which is not the (primary) focus of stylistic analysis. In particular, we wish to segment using information that goes beyond the distribution of words in the text being segmented. The model for stylistic segmentation we propose here is related to the TextTiling technique of Hearst (1994) and the style change function of Stamatatos (2009a), but our model is generalized so that it applies to any numeric metric (feature) that is defined over a span; importantly, style change curves represent the change of a set of very diverse features.

Our goal is to find the precise points in the text where a stylistic change (a voice switch)

occurs. To do this, we calculate, for each token in the text, a measure of stylistic change which corresponds to the distance of feature vectors derived from a fixed-length span on either side of that point. That is, if $\mathbf{v}_{ij}$ represents a feature vector derived from the tokens between (inclusive) indices i and j, then the stylistic change at point c_i for a span (window) of size w is:

$$c_i = Dist(\mathbf{v}_{(i-w)(i-1)}, \mathbf{v}_{i(i+w-1)})$$

This function is not defined within w words of the edge of the text, and we generally ignore the possibility of breaks within these (unreliable) spans. Possible distance metrics include cosine distance, euclidean distance, and city-block distance. In his study, Guthrie (2008) found best results with city-block distance, and that is what we will primarily use here. The feature vector can consist of any features that are defined over a span; one important step, however, is to normalize each feature (here, to a mean of 0 and a standard deviation of 1), so that different scaling of features does not result in particular features having an undue influence on the stylistic change metric. That is, if some feature is originally measured to be f_i in the span i to $i+w-1$, then its normalized version f'_i (included in $\mathbf{v}_{i(i+w-1)}$) is:

$$f'_i = \frac{f_i - \overline{f}}{\sigma_f}$$

The local maxima of c represent our best predictions for the stylistic breaks within a text. However, stylistic change curves are not well behaved; they may contain numerous spurious local maxima if a local maximum is defined simply as a higher value between two lower ones. We can narrow our definition, however, by requiring that the local maximum be maximal within some window w'. That is, our breakpoints are those points i where, for all points j in the span $x-w'$, $x+w'$, it is the case that $g_i > g_j$. As it happens, $w' = w/2$ is a fairly good choice for our purposes, creating spans no smaller than the smoothed window, though w' can be lowered

to increase breaks, or increased to limit them. The absolute height of the curve at each local minimum offers a secondary way of ranking (and eliminating) potential breakpoints, if more precision is required; however, in our task here the breaks are fairly regular but often subtle, so focusing only on the largest stylistic shifts is not necessarily desirable.

4.5.4 Features

The set of features we explore for this task falls roughly into two categories: surface and extrinsic. The distinction is not entirely clear cut, but we wish to distinguish features that use the basic properties of the words or their PoS, which have traditionally been the focus of automated stylistic analysis, from features which rely heavily on external lexical information, for instance word sentiment and, in particular, vector space representations, which are more novel for this task. First, the surface features:

Word length A common textual statistic in register and readability studies. Readability, in turn, has been used for plagiarism detection (Stein et al., 2011), and related metrics were consistently among the best for Guthrie (2008).

Syllable count Syllable count is reasonably good predictor of the difficulty of a vocabulary, and is used in some readability metrics.

Punctuation frequency The presence or absence of punctuation such as commas, colons, semicolons can be very good indicator of style. We also include periods, which offer a measure of sentence length.

Line breaks Our only poetry-specific feature; we count the number of times the end of a line appears in the span. More or fewer line breaks (that is, longer or shorter lines) can vary the rhythm of the text, and thus its overall feel.

Parts of speech Lexical categories can indicate, for instance, the degree of nominalization, which is a key stylistic variable (Biber, 1988). We collect statistics for the four main lexical categories (noun, verb, adjective, adverb) as well as prepositions, determiners, and proper nouns.

Pronouns We count the frequency of first-, second-, and third-person pronouns, which can indicate the interactiveness and narrative character of a text (Biber, 1988).

Verb tense Past tense is often preferred in narratives, whereas present tense can give a sense of immediacy.

Type-token ratio A standard measure of lexical diversity.

Lexical density Lexical density is the ratio of the count of tokens of the four substantive parts of speech to the count of all tokens.

Contextuality measure The contextuality measure of Heylighen and Dewaele (2002) is based on PoS tags (e.g. nouns decrease contextuality, while verbs increase it), and has been used to distinguish formality in collaboratively built encyclopedias (Emigh and Herring, 2005).

Dynamic In addition to the hand-picked features above, we test dynamically including words and character trigrams that are common in the text being analyzed, particularly those not evenly distributed throughout the text (we exclude punctuation). To measure the latter, we define *clumpiness* as the square root of the index of dispersion or variance-to-mean ratio (Cox and Lewis, 1966) of the (text-length) normalized differences between successive occurrences of a feature, including (importantly) the difference between the first index of the text and the first occurrence of the feature as well as the last occurrence and the last index; the measure varies

between 0 and 1, with 0 indicating perfectly even distribution. We test with the top n features based on the ranking of the product of the feature's frequency in the text (*tf*) or product of the frequency and its clumpiness (*tf-cl*); this is similar to a *tf·idf* weight.

Next, the extrinsic features. For those lexicons which include only lemmatized forms, the words are lemmatized before their values are retrieved.

Percent of words in Dale-Chall Word List A list of 3000 basic words that is used in the Dale-Chall Readability metric (Dale and Chall, 1995).

Average unigram count in 1T Corpus Another metric of whether a word is commonly used. We use the unigram counts in the 1T 5-gram Corpus (Brants and Franz, 2006). Here and below, if a word is not included it is given a zero.

Sentiment polarity The positive or negative stance of a span could be viewed as a stylistic variable. We test two lexicons, a hand-built lexicon for the SO-CAL sentiment analysis system which has shown superior performance in lexicon-based sentiment analysis (Taboada et al., 2011), and SentiWordNet (SWN), a high-coverage automatic lexicon built from WordNet (Baccianella et al., 2010). The polarity of each word over the span is averaged.

Sentiment extremity Both lexicons provide a measure of the degree to which a word is positive or negative. Instead of summing the sentiment scores, we sum their absolute values, to get a measure of how extreme (subjective) the span is.

Formality Average formality score, using a lexicon of formality (Brooke et al., 2010a) built using latent semantic analysis (LSA) (Landauer and Dumais, 1997).

Dynamic General Inquirer The General Inquirer dictionary (Stone et al., 1966), which was used for stylistic inconsistency detection by Guthrie (2008), includes 182 content analysis tags, many of which are relevant to style; we remove the two polarity tags already part of the SO-CAL dictionary, and select others dynamically using our *tf-cl* metric.

LSA vector features Earlier, in Section 3.4, we posited that, in highly diverse register/genre corpora, the lowest dimensions of word vectors derived using LSA (or other dimensionality reduction techniques) often reflect stylistic concerns, in particular finding that using the first 20 dimensions to build a formality lexicon provided the best results in a near-synonym evaluation. Here, we investigate using these LSA-derived vectors directly, with each of the first 20 dimensions corresponding to a separate feature. We test with vectors derived from the word-document matrix of the ICWSM 2009 blog dataset (Burton et al., 2009) which includes 1.3 billion tokens, and also from the BNC (Burnard, 2000), which is 100 million tokens. The length of the vector depends greatly on the frequency of the word; since this is being accounted for elsewhere, we normalize each vector to the unit circle.

4.5.5 Evaluation method

To evaluate our method we apply standard topic segmentation metrics, comparing the segmentation boundaries to a gold standard reference. The measure P_k, proposed by Beeferman et al. (1997), uses a probe window equal to half the average length of a segment; the window slides over the text, and counts the number of instances where a unit (in our case, a token) at one edge of the window was predicted to be in the same segment (according to the reference) as a unit at the other edge, but in fact is not; or was predicted not to be in the same segment, but in fact is. This count is normalized by the total number of tests to get a score between 0 and 1, with 0 being a perfect score (the lower, the better). Pevzner and Hearst (2002) criticize this

metric because it penalizes false positives and false negatives differently and sometimes fails to penalize false positives altogether; their metric, *WindowDiff* (WD), solves these problems by counting an error whenever there is a difference between the number of segments in the prediction as compared to the reference. Recent work in topic segmentation (Eisenstein and Barzilay, 2008) continues to use both metrics, so we also present both here.

During initial testing, we noted a fairly serious shortcoming with both these metrics: all else being equal, they will usually prefer a system which predicts fewer breaks; in fact, a system that predicts no breaks at all can score under 0.3 (a very competitive result both here and in topic segmentation), if the variation of the true segment size is reasonably high. This is problematic because we do not want to be trivially 'improving' simply by moving towards a model that is too cautious to guess anything at all. We therefore use a third metric, which we call BD (break difference), which sums all the distances, calculated as fractions of the entire text, between each true break and the nearest predicted break. This metric is also flawed, because it can be trivially made 0 (the best score) by guessing a break everywhere. However, the relative motion of the two kinds of metric provides insight into whether we are simply moving along a precision/recall curve, or actually improving overall segmentation.

We compare our method to the following baselines:

Random selection We randomly select boundaries, using the same number of boundaries in the reference. We use the average over 50 runs.

Evenly spaced We put boundaries at equally spaced points in the text, using the same number of boundaries as the reference.

Random feature We use our stylistic change curve method with a single feature which is created by assigning a uniform random value to each token and averaging across the span.

Again, we use the average score over 50 runs.

4.5.6 Artificial poems experiment

Our main interest is *The Waste Land*. It is, however, prudent to develop our method, i.e. conduct an initial investigation of our method, including parameters and features, using a separate corpus. We do this by building artificial mixed-style poems by combining stylistically distinct poems from different authors, as others have done with prose.

Our set of twelve poems used for this evaluation was selected by an English literature expert[11] to reflect the stylistic range and influences of poetry at the beginning of the twentieth century, and *The Waste Land* in particular. The titles were removed, and each poem was tagged by an automatic PoS tagger (Schmid, 1995). Koppel et al. built their composite version of two books of the Bible by choosing, at each step, a random span length (from a uniform distribution) to include from one of the two books being mixed, and then a span from the other, until all the text in both books had been included. Our method is similar, except that we first randomly select six poems to include in the particular mixed text, and at each step we randomly select one of the poems, reselecting if the poem has been used up or the remaining length is below our lower bound. For our first experiment, we set a lower bound of 100 tokens and an upper bound of 200 tokens for each span; although this gives a higher average span length than that of *The Waste Land*, our first goal is to test whether our method works in the (ideal) condition where the feature vectors at the breakpoint generally represent spans which are purely one poem or another for a reasonably high w (100). We create 50 texts using this method. In addition to testing each individual feature, we test several combinations of features (all features, all surface features, all extrinsic features), and present the best results for greedy feature removal, starting with all features (excluding dynamic ones) and choosing features to remove which minimize

[11] Adam Hammond, who was an author on the original paper this section is based on.

the sum of the three metrics.

The Feature Sets section of Table 4.8 gives the individual feature results for segmentation of the artificially-combined poems. Using any of the features alone is better than our baselines, though some of the metrics (in particular type-token ratio) are only a slight improvement. Line breaks are obviously quite useful in the context of poetry (though the WD score is high, suggesting a precision/recall trade-off), but so are more typical stylistic features such as the distribution of basic lexical categories and punctuation. The unigram count and formality score are otherwise the best two individual features. The sentiment-based features did more modestly, though the extremeness of polarity was useful when paired with the coverage of SentiWordNet. Among the larger feature sets, the GI was the least useful, though more effective than any of the individual features, while dynamic word and character trigrams did better, and the ICWSM LSA vectors better still; the difference in size between the ICWSM and BNC is obviously key to the performance difference here. In general using our *tf-cl* metric was better than *tf* alone.

When we combine the different feature types, we see that extrinsic features have a slight edge over the surface features, but the two do complement each other to some degree. Although the GI and dynamic feature sets do well individually, they do not combine well with other features in this unsupervised setting, and our best results do not include them. The greedy feature selector removed 4 LSA dimensions, type-token ratio, prepositions, second-person pronouns, adverbs, and verbs to get our best result. Our choice of w to be the largest fully-reliable size (100) seems to be a good one, as is our use of city-block distance rather than the alternatives. Overall, the metrics we are using for evaluation suggest that we are roughly halfway to perfect segmentation.

Table 4.8: Segmentation accuracy in artificial poems

Configuration	**Metrics**		
	WD	P_k	BD
Baselines			
Random breaks	0.532	0.465	0.465
Even spread	0.498	0.490	0.238
Random feature	0.507	0.494	0.212
Feature sets			
Word length	0.418	0.405	0.185
Syllable length	0.431	0.419	0.194
Punctuation	0.412	0.401	0.183
Line breaks	0.390	0.377	0.200
Lexical category	0.414	0.402	0.177
Pronouns	0.444	0.432	0.213
Verb tense	0.444	0.433	0.202
Lexical density	0.445	0.433	0.192
Contextuality	0.462	0.450	0.202
Type-Token ratio	0.494	0.481	0.204
Dynamic (*tf*, *n*=50)	0.399	0.386	0.161
Dynamic (*tf-cl*, 50)	0.385	0.373	0.168
Dynamic (*tf-cl*, 500)	0.337	0.323	0.165
Dynamic (*tf-cl*, 1000)	0.344	0.333	0.199
Dale-Chall	0.483	0.471	0.202
Count in 1T	0.424	0.414	0.193
Polarity (SO-CAL)	0.466	0.487	0.209
Polarity (SWN)	0.490	0.478	0.221
Extremity (SO-CAL)	0.450	0.438	0.199
Extremity (SWN)	0.426	0.415	0.182
Formality	0.409	0.397	0.184
All LSA (ICWSM)	0.319	0.307	0.134
All LSA (BNC)	0.364	0.352	0.159
GI (*tf*, *n*=5)	0.486	0.472	0.201
GI (*tf-cl*, 5)	0.449	0.438	0.196
GI (*tf-cl*, 50)	0.384	0.373	0.164
GI (*tf-cl*, 100)	0.388	0.376	0.163
Combinations			
Surface	0.316	0.304	0.150
Extrinsic	0.314	0.301	0.124
All	0.285	0.274	0.128
All w/o GI, dynamic	0.272	0.259	0.102
All greedy (Best)	**0.253**	**0.242**	**0.099**
Best, *w*=150	0.289	0.289	0.158
Best, *w*=50	0.338	0.321	0.109
Best, Diff=euclidean	0.258	0.247	0.102
Best, Diff=cosine	0.274	0.263	0.145

4.5.7 *The Waste Land* experiment

In order to evaluate our method on *The Waste Land*, we first created a gold standard voice switch segmentation. Our gold standard represents an amalgamation, by our English literature expert, of several sources of information. First, we enlisted a class of 140 undergraduates in an English literature course to segment the poem into voices based on their own intuitions, and we created a combined student version based on majority judgment. Second, our expert listened to the 6 readings of the poem included on *The Waste Land* app (Touch Press LLP, 2011), including two readings by T.S. Eliot, and noted places where the reader's voice seemed to change; these were combined to create a reader version. Finally, our expert amalgamated these two versions and incorporated insights from independent literary analysis to create a final gold standard.

We created two versions of the poem for evaluation: for both versions, we removed everything but the main body of the text (i.e. the prologue, dedication, title, and section titles), since these are not produced by voices in the poem. The 'full' version contains all the other text (a total of 68 voice switches), but our 'abridged' version involves removing all segments (and the corresponding voice switches, when appropriate) which are 20 or fewer tokens in length and/or which are in a language other than English, which reduces the number of voice switches to 28 (the token count is 3179). This version allows us to focus on the segmentation for which our method has a reasonable chance of succeeding and ignore the segmentation of non-English spans, which is relatively trivial yet potentially confounding. We use $w = 50$ for the full version, since there are almost twice as many breaks as in the abridged version (and our artificially generated texts).

Our results for *The Waste Land* are presented in Table 4.9. Notably, in this evaluation, we do not investigate the usefulness of individual features or attempt to fully optimize our solution using this text. Our goal is to see if a general stylistic segmentation system, developed

Table 4.9: Segmentation accuracy in *The Waste Land*

Configuration	**Metrics**		
	WD	P_k	BD
Full text			
Baselines			
Random breaks	0.517	0.459	0.480
Even spread	0.559	0.498	0.245
Random feature	0.529	0.478	0.314
System (w=50)			
Table 4.8 Best	0.458	0.401	0.264
GI	0.508	0.462	0.339
Dynamic	0.467	0.397	0.257
LSA (ICWSM)	0.462	0.399	0.280
All w/o GI	**0.448**	0.395	0.305
All w/o dynamic, GI	0.456	**0.394**	**0.228**
Abridged text			
Baselines			
Random breaks	0.524	0.478	0.448
Even spread	0.573	0.549	0.266
Random feature	0.525	0.505	0.298
System (w=100)			
Table 4.8 Best	0.370	0.341	0.250
GI	0.510	0.492	0.353
Dynamic	0.415	0.393	0.274
LSA (ICWSM)	0.411	0.390	0.272
All w/o GI	0.379	0.354	0.241
All w/o dynamic, GI	**0.345**	**0.311**	**0.208**

on artificial texts, can be applied successfully to the task of segmenting an actual stylistically diverse poem. The answer is yes. Although the task is clearly more difficult, the results for the system are well above the baseline, particularly for the abridged version. One thing to note is that using the features greedily selected for the artificial system (instead of just all features) appears to hinder, rather than help; this suggests a supervised approach might not be effective. The GI is too unreliable to be useful here, whereas the dynamic word and trigram features continue to do fairly well, but they do not improve the performance of the rest of the features combined. Once again the LSA features seem to play a central role in this success. We

manually compared predicted with real switches and found that there were several instances (corresponding to very clear voices switches in the text) which were nearly perfect. Moreover, the model did tend to predict more switches in sections with numerous real switches, though these predictions were often fewer than the gold standard and out of sync (because the sampling windows never consisted of a pure style).

4.6 Clustering voices of *The Waste Land*

4.6.1 Introduction

The work in the previous section focused on only the segmentation part of the voice identification task.[12] Here, we instead assume an initial segmentation and then try to create clusters corresponding to segments of the *The Waste Land* which are spoken by the same voice. Of particular interest is the influence of the initial segmentation on the success of this downstream task.

4.6.2 Method

Our approach to voice identification in *The Waste Land* consists first of identifying the boundaries of voice spans, as outlined in the previous section. Given a segmentation of the text, we consider each span as a data point in a clustering problem. The elements of the vector correspond to the best feature set from the segmentation task, with the rationale that features which were useful for detecting changes in style should also be useful for identifying stylistic similarities.

For clustering, we use a slightly modified version of the popular k-means algorithm (MacQueen, 1967). Briefly, k-means assigns points to a cluster based on their proximity to the k

[12]The work presented in this section is based on "Clustering voices in *The Waste Land*", by Julian Brooke, Graeme Hirst, and Adam Hammond, published in the *Proceedings of the 2nd Workshop on Computational Literature for Literature* (Brooke et al., 2013).

cluster centroids, which are initialized to randomly chosen points from the data and then iteratively refined until convergence, which in our case was defined as a change of less than 0.0001 in the position of each centroid during one iteration.[13] Our version of k-means is distinct in two ways: first, it uses a weighted centroid where the influence of each point is based on the token length of the underlying span, i.e. short (unreliable) spans which fall into the range of some centroid will have less effect on the location of the centroid than larger spans. Second, we use a city-block (L_1) distance function rather than standard Euclidean (L_2) distance function; in the segmentation task, we found that city-block (L_1) distance was preferred, a result which is in line with other work in stylistic inconsistency detection (Guthrie, 2008). Though it would be interesting to see if a good k could be estimated independently, for our purposes here we set k to be the known number of speakers in our gold standard.

4.6.3 Evaluation

We evaluate our clusters by comparing them to a gold standard annotation. There are various metrics for extrinsic cluster evaluation; Amigó et al. (2009) review various options and select the BCubed precision and recall metrics (Bagga and Baldwin, 1998) as having all of a set of key desirable properties. BCubed precision is a calculation of the fraction of item pairs in the same cluster which are also in the same category, whereas BCubed recall is the fraction of item pairs in the same category which are also in the same cluster. The harmonic mean of these two metrics is BCubed F-score. Typically, the 'items' are exactly what has been clustered, but this is problematic in our case, because we wish to compare methods which have different segmentations and thus the vectors that are being clustered are not directly comparable. Instead, we calculate the BCubed measures at the level of the token; that is, for the purposes of measuring performance we act as if we had clustered each token individually, instead of the spans of

[13] Occasionally, there was no convergence, at which point we halted the process arbitrarily after 100 iterations.

tokens actually used.

As in the preceding section, our first evaluation is against a set of 20 artificially-generated 'poems' which are actually randomly generated combinations of parts of 12 poems which were chosen (by an English literature expert, one of the authors) to represent the time period and influences of *The Waste Land*. The longest of these poems is 1291 tokens and the shortest is just 90 tokens (though 10 of the 12 have at least 300 tokens); the average length is 501 tokens. Our method for creating these poems is the same as in the preceding section. Again, the idea is to allow us to evaluate our method in more ideal circumstances i.e. when there are very distinct voices corresponding to different poets, and the voice spans tend to be fairly long.

Our gold standard annotation of *The Waste Land* speakers is far more tentative. As already mentioned, it is based on a number of sources: our own English literature expert, relevant literary analysis (Cooper, 1987), and also *The Waste Land* app (Touch Press LLP, 2011), which includes readings of the poem by various experts, including T.S. Eliot himself. However, there is inherently a great deal of subjectivity involved in literary annotation and, indeed, one of the potential benefits of our work is to find independent justification for a particular voice annotation. Our gold standard thus represents just one potential interpretation of the poem, rather than a true, unique gold standard. The average size of the 69 segments in the gold standard is 50 tokens; the range, however, is fairly wide: the longest is 373 tokens, while the shortest consists of a single token. Our annotation has 13 voices altogether.

We consider three segmentations: the segmentation of our gold standard (Gold), the segmentation predicted by our segmentation model (Automatic), and a segmentation which consists of equal-length spans (Even), with the same number of spans as in the gold standard. The Even segmentation should be viewed as the baseline for segmentation, and the Gold segmentation an "oracle" representing an upper bound on segmentation performance. For the automatic segmentation model, we use the settings from our work in the preceding section. We also

Table 4.10: Clustering results for artificial poems

Configuration	BCubed metrics		
	Prec.	**Rec.**	**F-score**
Initial Even	0.703	0.154	0.249
Initial Automatic	0.827	0.177	0.286
Initial Gold	1.000	0.319	0.465
Random Even	0.331	0.293	0.307
Random Automatic	0.352	0.311	0.327
Random Gold	0.436	0.430	0.436
k-means Even	0.462	0.409	0.430
k-means Automatic	0.532	0.479	0.499
k-means Gold	0.716	0.720	0.710
k-means Gold Seeded	0.869	0.848	0.855

compare three possible clusterings for each segmentation: no clustering at all (Initial), that is, we assume that each segment is a new voice; *k*-means clustering (*k*-means), as outlined above; and random clustering (Random), in which we randomly assign each voice to a cluster. For these latter two methods, which both have a random component, we averaged our metrics over 50 runs. Random and Initial are here, of course, to provide baselines for judging the effectiveness of *k*-means clustering model. Finally, when using the gold standard segmentation and *k*-means clustering, we included another oracle option (Seeded): instead of the standard *k*-means method of randomly choosing them from the available datapoints, each centroid is initialized to the longest instance of a different voice, essentially seeding each cluster.

4.6.4 Results

Table 4.10 contains the results for our first evaluation of voice clustering, the automatically-generated poems. In all the conditions, using the gold segmentation far outstrips the other two options. The automatic segmentation is consistently better than the evenly-spaced baseline, but the performance is actually worse than expected; the segmentation metrics we used in the preceding section suggested that the segmentation was roughly halfway to a perfect segmentation,

Table 4.11: Clustering results for *The Waste Land*

Configuration	**BCubed metrics**		
	Prec.	**Rec.**	**F-score**
Initial Even	0.792	0.069	0.128
Initial Automatic	0.798	0.084	0.152
Initial Gold	1.000	0.262	0.415
Random Even	0.243	0.146	0.183
Random Automatic	0.258	0.160	0.198
Random Gold	0.408	0.313	0.352
k-means Even	0.288	0.238	0.260
k-means Automatic	0.316	0.264	0.296
k-means Gold	0.430	0.502	0.461
k-means Gold Seeded	0.491	0.624	0.550

but the better segmentation is reflected mostly in precision and not recall; therefore clustering performance as expressed by the F-score is far less optimistic. Random clustering is clearly worse than *k*-means, but for the unreliable segmentations the harmonic mean is actually higher than the initial clustering, due to an increase in recall. The improvement due to *k*-means is sizable, and fairly consistent across the three segmentations, though better segmentations see more absolute improvement. Seeding is also quite effective, and for this relatively easy dataset we approach perfect performance under this condition.

The results for *The Waste Land* are in Table 4.11. Many of the basic patterns are the same, including the consistent ranking of the methods; overall, however, the clustering is far less effective. This is particularly true for the gold-standard condition, which only increases modestly between the initial and clustered state; the marked increase in recall is balanced by a major loss of precision. In fact, unlike with the artificial text, the most promising aspect of the clustering seems to be the fairly sizable boost to the quality of clusters in automatic segmenting performance. The effect of seeding is also very consistent, nearly as effective as in the automatic case.

We also looked at the results for individual speakers in *The Waste Land*; many of the speak-

ers (some of which appear only in a few lines) are very poorly distinguished, even with the gold-standard segmentation and seeding, but there are a few that cluster quite well; the best two are in fact our examples from the previous section[14] that is, the narrator (F-score 0.869), and the chatty woman (F-score 0.605). The former result is particularly important, from the perspective of literary analysis, since there are several passages which seem to be the main narrator (and our expert annotated them as such) but which are definitely open to interpretation.

4.6.5 Discussion

Literature, by its very nature, involves combining existing means of expression in surprising new ways, resisting supervised analysis methods that depend on assumptions of conformity. Our unsupervised approach to distinguishing voices in poetry offers this necessary flexibility, and indeed seems to work reasonably well in cases when the stylistic differences are clear. *The Waste Land*, however, is a very subtle text, and our results suggest that we are a long way from something that would be a considered a possible human interpretation. Nevertheless, applying quantitative methods to these kinds of texts can, for literary scholars, bridge the gap between abstract interpretations and the details of form and function (McKenna and Antonia, 2001). In our own case, this computational work is just one aspect of a larger project in literary analysis where the ultimate goal is not to mimic human behavior per se, but rather to better understand literary phenomena by annotation and modeling of these phenomena (Hammond, 2013; Hammond et al., 2013).

[14]These passages were selected by our expert for their distinctness, so the fact that they turned out to be the most easily clustered is actually a result of sorts (albeit an anecdotal one), suggesting that our clustering behavior does correspond somewhat to a human judgment of distinctness.

4.7 Intrinsic plagiarism detection at the PAN '12 shared task

4.7.1 Introduction

The task of intrinsic plagiarism detection involves distinguishing portions of a single text which are written by different authors (Stein et al., 2011).[15] Key characteristics of the task are the lack of texts written purely by one author or another, as is typically the case in authorship attribution (Stamatatos, 2009b), and lack of a database of texts from which the texts are formed, which is the focus of extrinsic plagiarism detection (Oberreuter et al., 2011). As such, it has more in common with (to the point of being arguably synonymous with) the task of stylistic inconsistency detection (Graham et al., 2005; Guthrie, 2008; Koppel et al., 2011), and our approach in the task is strongly influenced by this work.

Relatively successful approaches to intrinsic plagiarism detection (Stamatatos, 2009a; Oberreuter et al., 2011; Kestemont et al., 2011) have often relied exclusively on variation in word or character *n*-gram frequency as the key indicator of stylistic variation, an approach that is clearly effective in the general task of authorship attribution (Stamatatos, 2009b). However, in the context of spans as small as a paragraph, we are somewhat skeptical that these sorts of features capture anything much beyond the topic shifts which are a common artifact of the usually artificially-created test sets. In fact, in the context of another stylistic text classification task, native language identification (see Chapter 5), we found evidence that the effectiveness of character *n*-grams as stylistic features seemed to derive largely from the confounding effects of topic in the corpus (Brooke and Hirst, 2011); when topic was (partially) controlled for, performance of these features plummeted by over 30%. In the case of real-world plagiarism, we would expect that differences in style, not topic, would be the key indicator of plagiarism, and, although focusing on surface (intrinsic) features may provide superficial improvement in arti-

[15] This work was adapted from "Paragraph clustering for intrinsic plagiarism detection using a stylistic vector-space model with extrinsic features" by Julian Brooke and Graeme Hirst, published in the *Notebook for PAN 2012 Lab at CLEF* (Brooke and Hirst, 2012c).

ficial settings, we think it is important to branch out and incorporate stylistic information that reflects underlying dimensions of stylistic variation; a similar approach applied to various text classification tasks has shown promise (Argamon et al., 2007). Stylistic segmentation of a real stylistically diverse document, *The Waste Land* (Section 4.5), we compared the typical surface *n*-gram features to richer extrinsic features, and found that the *n*-gram surface features, though reasonably useful on their own, did not seem to combine well with more-targeted features, and we ultimately discarded them. Therefore, in the present work, which is in most other respects a reasonably straightforward clustering approach based on maximizing vector distance between author spans, we entirely eschew *n*-gram features in favor of the linguistically motivated extrinsic features that we applied to poetry segmentation. In addition, we use a novel approach based on modeling expected random differences to attenuate the effects of variation in span length.

The intrinsic plagiarism portion of the PAN '12 task contained two subtasks, both of which involved clustering paragraphs in a small set of texts. In the first, the 'plagiarism' was one-time, consecutive paragraph 'intrusion' of a single author into the text of a different author; the total number of paragraphs in each text is 20, and it was also possible that there was no intrusion at all. Our result on this first task was decent, 88.8% correct (which put us roughly in the middle of those who participated). The second, multi-author task involved texts which were a mixture of 2 to 4 authors, with no information about the ordering or number of paragraphs per author, for a total of 30 paragraphs. Our results on this task were poor (45.6%), tied for worst among the submissions, for various reasons that will be addressed in the discussion.

4.7.2 Feature Selection and Extraction

The set of features that we explore for this task is the same as for the segmentation of *The Waste Land* in Section 4.5, with the exception that we do not use the line break feature since it

does not apply outside the poetry domain.

4.7.3 Clustering

Our general approach to both of the paragraph-clustering subtasks of the intrinsic plagiarism detection task is to assign paragraphs into author groups that maximize the (average) distance between authors. Following Guthrie's work in stylistic outlier detection (Guthrie, 2008) and our own previous conclusions (Section 4.5), we use L_1 or city block distance as the distance metric. Another important insight of Guthrie that is it is desirable to use spans as large as possible, i.e. we consider the distance between the spans suspected to be written by a single author, rather than the distances between individual paragraphs (e.g. a graph-based approach). In particular, for the single-intruder task, we considered all possible start/end pairs for second author intrusion, and calculated the difference between the main and intruder spans, choosing the pair that produced a maximal distance. For the multi-author task, we began by assigning all paragraphs to a single author and none to the three other authors. We then iteratively moved spans from one author group to another, each step being the one that provided the maximum increase in average distance, until no further improvement was possible.

However, there is a serious flaw in this kind of approach: all other things being equal, shorter spans have more random variation and thus are, on average, more distant (sometimes *much* more distant) from any given span than a longer, more homogeneous span. Fortunately, this effect can be modeled. We did this by calculating the expected distance of sums of random variables. Supposing a span of some basic length (we used 50 tokens) to have a random component of normal distribution (with a mean and standard deviation of 1), we can estimate the expected influence of this randomness on the distance measure between any pair of spans — for instance, spans of length 400 and 100 — by looking at the sum of random variables corresponding to the n basic spans that make it up — in this example, the expected difference

between the sum of 8 random variables and the sum of 2 random variables with the same distribution. We ran 100 trials using a random number generator and computed a table of such expected differences, and then divided our calculated distance by the corresponding number in the table to get a new distance that takes expected difference into account.[16]

4.7.4 Evaluation

Even before we reached the final version described above, our approach had perfect performance on the original example texts provided by the PAN organizers, so we created some additional corpora for testing, collecting a few different types of texts (early modern novels, translated Russian novels, and political treatises) from Project Gutenberg, and automatically creating mixed texts of various difficulty. Here, we present results using two relatively easy corpora which consist of texts of paragraphs randomly pulled from novels by Fyodor Dostoyevsky (in English translation) and Thomas Hardy, and political texts by Thomas Paine and Jean-Jacques Rousseau. For each text in the corpus for the mixed-author task, we first choose the number of authors (between 2 and 4), then randomly selected the authors, the number of paragraphs (between 10 and 30) and then the paragraphs themselves from random locations in the text. For the two-author insertion task, we randomly choose two authors, a total number of paragraphs, and two non-equal indices within that range; for each author, a random starting location was randomly selected and consecutive paragraphs from the first author were randomly selected for paragraphs before the first index, and then after the second, and consecutive paragraphs from the second author were inserted between the two indices. Both corpora have 30 texts created in this fashion.

There are various metrics for extrinsic cluster evaluation; Amigó et al. (2009) review various options and select the BCubed precision and recall metrics (Bagga and Baldwin, 1998) as

[16]There may be a closed-form solution to this problem, but in our case it was easier to derive it empirically.

Table 4.12: Clustering results with BCubed metrics on our test data.

Distance calculation	**Multi-author**			**Insertion**		
	Prec.	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**
Random baseline	0.411	0.378	0.386	0.754	0.694	0.704
Individual paragraph	**0.589**	0.620	0.582	0.818	**0.969**	**0.879**
Combined span	0.426	0.794	0.543	0.749	0.923	0.818
Combined span w/expected adjustment	0.533	**0.871**	**0.645**	**0.905**	0.866	**0.879**

having all of a set of key desirable properties. We discussed these metrics earlier in section 4.6.

We compare our algorithms with task-specific random baselines (with 50 trials) and two related alternatives: one which excludes our expected difference corrector and another that is based purely on the maximizing distances between individual paragraphs in the spans, rather than treating each cluster as a whole. The results for each of the two tasks are in Table 4.12.

There is little doubt that our expected difference adjustment has an overall positive effect, and, in the multi-author task, this provides the best result by a reasonably large margin. For the insertion task, which has a much higher random baseline, the individual span distance comparison was found to be roughly equivalent to our combined span approach with the adjustment.

4.7.5 Discussion

Our linguistically motivated, vector-space clustering approach shows promise, particularly with our expected difference adjustment. There is, however, obviously more work to do in this regard; for instance, using this adjustment our multi-author method never, in practice, predicts more than two authors, probably because the differences between short spans are now being underestimated rather than overestimated, meaning that two relatively short author spans (e.g. 3rd and 4th authors) are now highly dispreferred under our distance-maximizing algorithm. This may partially explain our relatively poor performance on the multi-author intrinsic plagiarism task, but in fact there is a more obvious reason. For instance, here are two paragraphs

from different authors in the multi-author task evaluation data (text1):

> John did not dream about the deli. He had nightmares of Douglas falling onto swords of knights on horseback, and woke several times throughout the night sweating and breathing heavily.
>
> But in an empty house, surrounded by evidence of Caroline's long absence, Hillie's words plagued him, and he was forced to accept that his mind might be capable of the cruelest of tricks. He felt desperately, hopelessly alone.

Stylistically, we find the two authors nearly indistinguishable. There are small differences (the second author prefers longer sentences and hyperbole), but the easiest way, for either human or computer, to identify the two is by the names of the characters. All but two of the paragraphs contains a proper name that appears in several other paragraphs (one author talks about Geoff, Hillie, and Caroline, the other about Douglas, John, and Mrs. Cumberland). Beyond proper names, there are also recurring topics: mail in one story, a job at a deli in another.[17] Any model that uses word or character *n*-grams should be able to take easy advantage of these regularities. Our model, conversely, was specifically designed not to do so; rather, it was developed to detect significant stylistic differences. In fact, in the task evaluation data there are quite clearly more obvious stylistic differences between some excerpts from the same novel than between some excerpts from different novels:

> On his departure, Hillie had pressed her business card into his hand. "My number's on there," she told him. "Call me, all right? I want you to promise." "I'm sorry," Geoff said. "I don't see the point." "You've suffered a shock. You can't be expected to cope at home on your own." Geoff had simply smiled at her. "I won't be on my own," he said. "I keep telling you. Caroline will look after me."

[17] There are even two sets of repeated paragraphs in this particular text: paragraphs 14 and 21 are the same, as are paragraphs 24 and 30!

Because of the presence of dialogue, this passage is radically different, stylistically, from the other passage from the same novel that was shown above. However, it is clear which of the two novels it comes from, since there are several proper-noun indicators. Given the range of subgenres within the novel genre, i.e. narration, description, and dialogue, this genre is a particularly bad choice for the purposes of simulating intrinsic plagiarism, since those stylistic features which exist and might be useful for distinguishing authors will be ultimately be drowned out by this confounding variation. Instead, topic-related features, which would be highly unreliable in the real world (for reasons that are obvious; what is the purpose of a student plagiarizing something that is topically distinct from the matrix text in which it is embedded?), are strongly preferred. Thus, we question whether the PAN evaluation is a useful reflection of the real-world task of intrinsic plagiarism detection: in particular, the evaluation should to be constructed so that the focal task is not confounded by orthogonal issues such as subgenre detection.

4.8 Style and discourse in *To the Lighthouse*

Our second literature project, "The Brown Stocking", focuses on a literary text chosen for its deliberately ambiguous nature: Virginia Woolf's (1927) *To the Lighthouse* (*TTL*).[18] There are two principal distinguishing features in Woolf's narrative style. The first is the tendency to reflect incidents through the subjective perspectives of characters rather than presenting them from the objective viewpoint of the narrator; thus *TTL* becomes a work where there is often more than one interpretation. Woolf's technique not only introduces multiple interpretations, however, but also blurs the transitions between individual perspectives, making it difficult to know in many instances who is speaking or thinking.

[18] A portion of this section was included in "A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together" by Adam Hammond, Julian Brooke, and Graeme Hirst, published in the *Proceedings of the 2nd Workshop on Computational Literature for Literature* (Hammond et al., 2013).

Woolf achieves this effect—multiple subjective impressions combined with obscuring of the lines separating them from the narrator and from one another—chiefly through the narrative technique of free indirect discourse (also known as free indirect style). Whereas direct discourse reports the actual words or thoughts of a character, and indirect discourse summarizes the thoughts or words of a character in the words of the entity reporting them, free indirect discourse (FID) is a mixture of narrative and direct discourse (Abrams, 1999). As in indirect discourse, the narrator employs third-person pronouns, but unlike indirect discourse, the narrator includes words and expressions that indicate subjective or personalized aspects clearly distinct from the narrator's style. Below are the opening sentences of *TTL*:

> "Yes, of course, if it's fine tomorrow," said Mrs. Ramsay. "But you'll have to be up with the lark," she added. To her son these words conveyed an extraordinary joy, as if it were settled, the expedition were bound to take place, and the wonder to which he had looked forward, for years and years it seemed, was, after a night's darkness and a day's sail, within touch.

Here, we are presented with two spans of objective narration (*said Mrs. Ramsay* and *she added*) and two passages of direct discourse, in which the narrator introduces the actual words of Mrs. Ramsay (*"Yes, of course, if it's fine tomorrow"* and *"But you'll have to be up with the lark"*). The rest of the passage is presented in FID, mixing together the voices of the narrator, Mrs. Ramsay, and her son James: while the use of third-person pronouns and the past tense clearly indicates the voice of the narrator, phrases such as *for years and years it seemed* clearly present a subjective perspective.

To explore this phenomenon, we have collected two rounds of student annotation. As part of a class project, we twice had 160 students mark up passages of between 100–300 words in accordance with Text Encoding Interface (TEI) guidelines. Students were instructed to use the TEI `said` element to enclose any instance of character speech (those which were not identified as speech were assumed to be narration), to identify the character whose speech is being

introduced, and to classify each of these instances as either direct, indirect, or free indirect discourse, and as either spoken aloud or thought silently. Because there are often several valid ways of interpreting a given passage, and because we are interested in how different students respond to the same passage, each word span was assigned to three or four students. This first round of annotation focused only on the first four chapters of *TTL*; raw average agreement of the various annotations at the level of the word was slightly less than 70%;[19] given the highly subjective nature of the task, levels of agreement typically required are likely to be beyond our reach. The second round of annotation involved the last seven chapters of the novel: the guidelines were tweaked and improved, but the basic annotation schema remained the same.

Since the annotation was carried out as an assignment, the students were strongly motivated to produce reasonable annotations. Nevertheless, a potential criticism of this work is our reliance on annotations from non-experts, where the subjectivity of the task makes it difficult to measure the extent to which an annotation might be mere nonsense rather than a reasonable alternative interpretation. A manual review of the annotations did reveal some consistent errors, though not enough for us to conclude that the annotations as a whole were compromised, and so we use them as is.

Though it is a long-term interest of the project, we do not focus on building a full model of variation in *TTL* here: more so than in *The Waste Land* (which moves almost erratically from character to character), to follow this text in terms of character shifts would require sophisticated models of discourse, anaphora resolution, etc., ones that are trained for novels in general and (less coherent) modernist novels in particular. Instead, like our exploration of lexical sociolinguistics earlier in this chapter, we will simply demonstrate that automatically-derived, large-coverage stylistic lexicons might play a role in such a model. In this experiment, we will

[19] Since each passage was tagged by a different set of students, we cannot apply traditional kappa measures. Raw agreement overestimates success, since unlike kappa it does not discount random agreement, which in this case varies widely across the different kinds of annotation.

use the 6-style lexicon from Section 3.7 to see how these styles are reflected in different kinds of discourse.

We will divide up the analysis by the two annotations, for various reasons: first, conclusions reached on two independent samples (here, collected from two different classes) are likely to be more reliable. Second, there are important differences between these two annotations in terms of the characters who are involved: the beginning of the novel is dominated by older characters such Mrs. Ramsay and Charles Tansley, while by the end of the novel younger characters such as the maid Lily and Ramsay children James and Cam are much more prominent. Our focus is the differences across types of discourse, but the different characters might have their own stylistic effects (though not, perhaps, to the extreme we saw in *The Waste Land*). We will limit our analysis to four major discourse types: narration, spoken direct speech, silent direct speech, and free indirect discourse. Indirect speech is fairly rare in the novel, and a manual analysis found that many annotations that used it were clear errors, so it is excluded.

We extracted all the words from the annotation parts of *TTL*, and built a stylistic lexicon using the highest performing LSA method from our earlier work (Section 3.7). For each style, we normalized the words in the dictionary to a mean of 0 and standard deviation of 1. To decide the correct discourse annotation for any given word, we took the majority annotation;[20] when there was no majority the annotation was discarded. For our results here, we excluded common words, defined as words that appeared at least 10 times in the text, which included common function words, names (which were excluded from the lexicon anyway), and very common verbs like *said* and *thought*. For each discourse annotation, the stylistic vectors for all the words were averaged. The results for the first annotation are in Table 4.13.

The most striking distinction in this table is between silent and spoken direct speech. In this first annotation, silent speech is much more abstract, and much less concrete, than all

[20] In doing so, we are minimizing one of our major interests in this student dataset, namely looking at disagreement as not merely error but rather (potentially) alternative but valid interpretations (Hammond et al., 2013).

Table 4.13: Average styles for various discourse types in Part 1, Chapters 1–4 of *To The Lighthouse*.

Discourse Type	Styles					
	Literary	**Abstract**	**Objective**	**Colloquial**	**Concrete**	**Subjective**
Narration	−0.26	−0.16	−0.22	+0.45	+0.12	−0.03
FID	−0.21	−0.02	−0.18	+0.42	−0.03	+0.14
Direct, Aloud	−0.33	−0.06	−0.43	+0.89	−0.09	+0.28
Direct, Silent	+0.13	+0.53	+0.03	+0.49	−0.56	−0.15

Table 4.14: Average styles for various discourse types in Part 3, Chapters 7–13 of *To The Lighthouse*.

Discourse Type	Styles					
	Literary	**Abstract**	**Objective**	**Colloquial**	**Concrete**	**Subjective**
Narration	−0.26	−0.33	−0.28	+0.32	+0.30	−0.05
FID	−0.18	−0.18	−0.21	+0.31	+0.14	+0.00
Direct, Aloud	−0.36	−0.11	−0.34	+0.80	+0.02	+0.05
Direct, Silent	−0.38	+0.03	−0.28	+0.80	−0.12	+0.15

other types of discourse. This makes perfect sense (given that silent speech is inherently an internal event); unfortunately, the effect is so strong that it pulls other styles along with it (along oral/written lines). Given its definition, we would expect that FID would fall between the extremes of spoken direct speech and narration, and for several key dimensions this is exactly the pattern: for example, spoken direct speech is more subjective, narration is less subjective, and FID is in the middle; narration is more concrete, spoken direct speech is less concrete, and FID is in the middle. But FID in *TTL* is consistently more 'written' (situational) than either narration or spoken direct speech; like silent direct speech here, it tends towards the abstract (though not to the same extreme).

Table 4.14 shows the results for the second round of annotation. Though the numbers have shifted slightly, the basic story at the oral pole remains the same, with narration being concrete, direct speech being subjective, and FID falling between the two. The major difference is in

silent direct speech, which still tends toward abstractness, but is otherwise far less extreme. The best explanation for this is that the internal speech is coming from the heads of much less introspective, self-serious characters, and and in general now the silent speech is much more in line with the spoken. Interestingly, the FID and even the narration is less abstract than in the earlier chapters. Otherwise, though, the FID is still more objective (formal) and more literary than either direct speech or narrative. In *TTL*, the free indirect discourse represents the 'meat' of the novel, where Woolf is free to use language that would not be spoken aloud by a character, or put forward by an disinterested narrator.

In this small study, we have shown how our stylistic dimensions are related to sub-genres in the novel, *To The Lighthouse*. The more predictable results, e.g. the distribution of subjectivity and concreteness, increased abstractness in silent speech, show that our stylistic dimensions are correctly reflecting important aspects of these subgenres, even in cases where larger correlations could have interfered. The fact that FID doesn't appear to be simply an mixture of speech and narration when more aesthetic styles are considered, though, is an novel point with respect to literary research on this phenomenon, and worthy of further study from that perspective (Adam Hammond, personal communication). In either case, this information is likely to be useful in future modeling of this phenomenon, especially since it is coming from words that are otherwise rare in this text, and thus would not be covered in an approach based on purely functional distinctions.

Chapter 5

Native Language Identification

Native language identification (NLI) is the task of identifying an author's native language (L1) based on a sample of second language (L2) writing. It should be clear that this task has a strong stylistic component, but it is fairly distinct from the other stylistic tasks we have already looked at; each L1 has its own unique influence on the L2 writing, so our polar conception of style breaks down to some extent. In response to a serious problem with previous work which relied on cross-validation in small corpora with significant topic biases, I adopt a cross-corpus approach to the task, showing that lexical features play an important role when sufficient data is available. A summary of my contributions in this area is given in Table 5.1.

5.1 Related work

The earliest focused work on native language detection was by Koppel et al. (2005). They classified texts from the International Corpus of Learner English (ICLE) into one of five (European) native language backgrounds using support vector machines. They described their feature set as stylistic; features included the frequency of function words, rare POS bigrams, letter *n*-grams, and spelling errors. They reported a performance of just over 80% accuracy on the task using the full feature set. Other early work on the ICLE includes that of Tsur and Rappoport (2007), who were concerned with identifying phonological language transfer;

Table 5.1: Overview of contributions in Chapter 5, including the focus of relevant projects, methods used, and conclusions reached.

- Section 5.2

  Focus Corpus building

  Methods Web crawling

  Conclusions Built Lang-8, a large multi-L1 corpus

- Section 5.3

  Focus NLI using L1 corpora

  Methods Collected, processed L1 corpora, calculating L1-influence using translated *n*-grams, unsupervised classification

  Conclusions Information from L1s can be used directly for NLI, good for Asian languages

- Section 5.4

  Focus Cross-corpus NLI

  Methods Supervised classification, testing major classifier options, domain adaptation by bias shift, testing major feature options

  Conclusions Cross-corpus NLI works well, but domain adaptation helps, lexical features are key, Lang-8 is robust for training, ICLE is problematic for training

- Section 5.5

  Focus NLI shared task

  Methods Feature testing, combining multiple corpora, domain adaptation, building robust models, using extra web data

  Conclusions Failure to find new useful features, domain adaptation allows for improvement with new corpora, success creating robust model without data from testing corpus

- Section 5.6

  Focus Identifying effect of corpus variables

  Methods Cross-corpus classification, proxy metrics, 6 different training and testing corpora

  Conclusions All variables affect classification to some degree, no single best training corpus, proficiency and topic variation across L1 important, metrics useful but overlapping

they focused on the construction of character *n*-gram models, reporting 66% accuracy with just these sub-word features, with only a small drop in performance when the dominant topic words in each sub-corpus (as identified using *tf·idf* were removed. Wong and Dras (2009) investigated particular types of syntactic error: subject-verb disagreement, noun-number disagreement, and determiner problems, relating the appearance of these errors to the features of

relevant L1s. However, they reported that these features do not help with classification, and they also note that character *n*-grams, though effective on their own, are not particularly useful in combination with other features.

In follow-up work, Wong and Dras (2011) attained 80% performance on a 7-language task using syntactic CFG production rules. Recent work by Wong et al. (2012) and Swanson and Charniak (2012) has explored the use of statistical grammatical induction techniques—Adaptor Grammars in the former case, Tree Substitution Grammars in the latter—to select better syntatic features for classification. Another interesting idea is the use of cohesion and word sophistication metrics (Crossley and McNamara, 2012). Traditionally, lexical features have been avoided when working in the ICLE, due to topic bias (see discussion of corpora in the next section), but some very recent work in the corpus has nevertheless focused on word *n*-grams (Bykh and Meurers, 2012), reaching 7-way performance scores of over 90%.

The work of Kochmar (2011) is distinct from those above in a number of ways: she used a different corpus of essays, derived from the Cambridge Learner Corpus[1], and concentrated on pairwise (SVM) classification within two European language sub-families. An exhaustive feature analysis indicated that character *n*-gram frequency is the most useful feature type for her task; unlike Wong and Dras (2011), syntactic production rules provided little benefit. With respect to lexical features, Kochmar presented some results using word *n*-grams, but regarded them as attributable to topic bias in the corpus. Error-type features (e.g. spelling, missing determiner) as provided by the corpus annotation offered little improvement over the high performance offered by the distributional features (e.g. POS/character *n*-grams).

Golcher and Reznicek (2011) used a string-distance metric to identify the native language of German learners in the Falko corpus (Lüdeling et al., 2008), and contrasted this with a topic classification task in the same corpus. Even after taking steps to mitigate topic bias

[1] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus

(removing the influence of the words in the title), the usefulness of the three feature types that they investigated (word token, word lemma, and POS) was remarkably similar across the two tasks, with the word features dominating in both cases. Surprisingly, the effect of POS was higher in topic classification than it was on L1-classification.

Finally, we note that native language identification has also been included as an element of larger author profiling studies (Estival et al., 2007; Garera and Yarowsky, 2009). A closely related task is the identification of translated texts and/or their language of origin (Baroni and Bernardini, 2006; van Halteren, 2008; Koppel and Ordan, 2011), though the tasks are distinct because the learners included in native language identification studies are usually at a level of linguistic proficiency below that of a professional translator (who in any case may be writing in his or her L1, rather than an L2) and are not operating under the requirement of faithfulness to some original text. Distinguishing whether or not a text is non-native (Tomokiyo and Jones, 2001) is also a related task, but most work in the area of L1 identification, including ours, assumes that we already know that a text was produced by a non-native speaker.

One potential application of NLI is in author profiling, which can be used to identify those who misrepresent themselves online (Fette et al., 2007). Another important use is as a pre-processing step to ESL error correction (Leacock et al., 2010): for example, Rozovskaya and Roth (2011) use L1-specific information to improve their preposition-correction system, while recent work in collocation correction relies on the specific forms present in the native language (Chang et al., 2008b; Dahlmeier and Ng, 2011).

5.2 Multi-L1 learner corpora

Until very recently, all nearly all work in NLI was done in the International Corpus of Learner English (Granger et al., 2009), which in its current version contains 6,085 essays from 16 different languages. This corpus is intended to reflect, among other things, the state of EFL

teaching in each of the countries around the world. An obvious challenge in building a corpus like the ICLE is incorporating the work of many researchers, educators and, of course, learners from different countries into a coherent whole. An original list of topics was chosen by the coordinating team, but leeway was clearly given to the organizers in each country, since some of the topics were, for instance, only relevant to Europeans (e.g. the future of a united Europe). Even when the original topic list was used, there were obvious biases in the particular topics chosen, with certain L1 backgrounds being dominated by certain topics. This explains why many NLI researchers have avoided word features when working with the ICLE; a classifier can simply learn to distinguish L1 by distinguishing topics. However, the problem extends deeper than that: we believe that certain topics are correlated with entirely different registers, which might have an effect on features that go beyond topic words. For example, many of the most common topics in the French subset of the corpus involve the relatively esoteric subjects of literature, religion, and politics, which might be discussed in a fairly formal register. In the Japanese corpus, however, we found a number of topics that were far more personal, for instance experience as an English learner and favorite travel destinations, which would likely be expressed in a more narrative and more colloquial manner. Arguably, these might reflect real differences in culture, but in the context of a corpus that cannot possibly reflect the full range of genres, we believe that these variations are extremely confounding for machine-learning based NLI, and they can affect a full range of feature types. We provide quantitative evidence of this problem later in Section 5.4.5.

Before we move on to newer, less problematic corpora, we will briefly consider one alternative, which has been recently proposed by Jarvis and Paquot (2012): filtering the ICLE at both the text level and the *n*-gram level to produce an unbiased corpus. Jarvis and Paquot suggest removing all texts from learners from Chinese, Japanese, Turkish, and Tswana backgrounds, since these have considerable variation from the others in terms of both topic and competency.

However, these four groups share another important distinction: They represent all the non-European L1s in the corpus. That means in order to minimize these confounding effects, we would have to limit ourselves entirely to European languages, an entirely unacceptable compromise, since the properties of language transfer within closely related languages is likely to be entirely different from those between families; for example, Europeans may struggle with spelling errors between numerous close cognates, but this is not an issue for a Chinese speaker, who must instead contend with various lexical bundles that are directly translated across European languages but have no exact equivalent in Chinese. For native language identification as a real world task, a full range of languages must be considered. More generally, controlling for competency is a complicated problem because distance between L1 and L2 is likely to be a huge determining factor in competency; it is very difficult to separate the two and, if the goal is to improve performance of an algorithm for NLI, it is not clear that learner proficiency should be controlled for at all. Moreover, Jarvis and Paquot removed *n*-grams that appeared both in prompts as well as commonly in the learner texts. Though this certainly would help remove some of the topic bias, the examples they provide demonstrate the limitations of this approach: from one text, they remove *society* and *prison*, but preserve other topical words such as *punish*, *criminal*, and *rehabilitate*, which are just as problematic. Presumably, one could push this further, removing more and more words, but we predict that this would almost immediately impinge on true L1 transfer features (for instance, preferring a close cognate), undermining the ultimate goal of NLI. This approach can certainly be applied to improve the reliability of relevant language-transfer research, which is Jarvis and Paquot's interest, but, again, if the ultimate goal of the research is developing robust high-performing NLI systems, discarding L1s and key features is not, we believe, a good way to begin.

Instead, we built a much larger but messier corpus from the web.[2] The Lang-8 website[3] provides a means for language learners to practice by writing journal entries in the language they are studying, which in turn is corrected by native speakers of that language who visit the site. We extracted a large collection of journals from the site, including 154,702 entries, or 22 million words. The site is based in Japan, and so learners of East Asian origin are disproportionately represented;[4] however, among the entries in our corpus there are 65 different native languages included, with 14 of those languages having at least 1,000 entries. Compared to the numerous variables that are recorded in manually collected learner corpora such as the ICLE, the information we have about each entry is rather minimal: other than (self-reported) native language and target language, we have a (unique) user name and the time which the entry was posted, though we use neither in the investigations reported here. There is some additional information available in the user profiles (e.g. gender), but we did not collect this information.

The ICLE contains primarily argumentative essays. The Lang-8 journal entries, by contrast, tend to be short personal narratives, though there are many exceptions: some users post their homework assignments, or ask for explicit translation or correction of a particular phrase out of the context of a coherent discourse. Though we did not carry out a rigorous analysis, the overall quality of the Lang-8 entries, i.e. the English proficiency of the users, seems to be generally much lower than the ICLE texts (which are written by university students). Moreover, the Lang-8 texts, because they are written entirely at the discretion of the user, appear to be more error-avoiding (Corder 1974); for the most part, users stay in their comfort zones, an

[2]This corpus was originally introduced in "Native Language Detection with 'Cheap' Learner Corpora" by Julian Brooke and Graeme Hirst, presented at the 2011 Conference of Learner Corpus Research (Brooke and Hirst, 2011)

[3]http://lang-8.com/

[4]The token counts for the best represented L1s in the Lang-8 corpus, in millions of tokens, are as follows: Japanese, 7.79; Chinese (both Mandarin and Cantonese), 5.66; Korean, 4.31; Russian, 1.00; Spanish, 0.52; French, 0.39; German, 0.26; Polish, 0.25; Italian, 0.23; Vietnamese, 0.20; Indonesian, 0.20; Arabic, 0.19; Portuguese, 0.16; Thai,0.15. All other L1s have fewer than 100,000 tokens

effect which we posit is amplified by the knowledge that their text may be critiqued by native speakers. On the other hand, there are (presumably) no limits on the time or other resources that users may use to create the entries, so some entries may represent a fairly major investment, including revisions.

A third learner corpus we will use throughout this section is a small sample of the First Certificate in English (FCE) portion of the Cambridge Learner Corpus, which was released for the purposes of essay scoring evaluation (Yannakoudakis et al., 2011); 16 different L1 backgrounds are represented. Each of the 1244 texts consists of two short answers in the form of a letter, a report, an article, or a short story, each tagged with the score provided by a trained examiner. The texts are also marked for specific usage errors, though we stripped this information in our pre-processing step.

A few other corpora have recently become available; they were not included in our original work in NLI (and indeed would complicate things due to different L1s), but we will present results using them in Sections 5.5 and 5.6. Perhaps most notable is the TOEFL-11 corpus, which was built by ETS specifically for NLI (Blanchard et al., 2013), and used for the 2013 NLI shared task (Tetreault et al., 2013).[5] It contains college TOEFL essays for 11 L1s, 1100 texts per L1, across 8 topics. It is also annotated for 3 difficultly levels. The International Corpus Network of Asian Learners of English or ICNALE (Ishikawa, 2011) is a collection of essays from college students in 10 Asian countries, with some proficiency information. There are only two topics in the corpus. Finally, like most of the other corpora, the International Corpus of Crosslinguistic Interlanguage (Tono et al., 2012) is also an essay corpus, though in contrast with other corpora it is focused on young learners, i.e. those in grade school. It includes both descriptive and argumentative essays on a number of topics, though these topics are scattered somewhat haphazardly across L1s.

[5]Though it was made available temporarily via a non-disclosure agreement for the purposes of the shared task, it is not officially available as of this writing.

5.3 Deriving lexical information for NLI from L1 texts

5.3.1 Introduction

In Second Language Acquisition (SLA) research, an *interlanguage*[6] is an emerging second language (L2) system (Selinker, 1992).[7] One of the defining qualities of an interlanguage is the use of native language (L1) features, a phenomenon which is known more generally as *language transfer* (Odlin, 1989). Though in related languages this may provide a early boost to learning, *language interference* is often the result where the two systems differ significantly, with learners continuing to use L1 features that are not appropriate to the L2, even after years of exposure.

Most previous work in L1 identification has avoided standard lexical features (e.g. word *n*-grams); the reason for this is not that these features would not be useful, but rather that there is significant topic variation across the languages in the corpora used for this task. Our work on the ICLE (see Section 5.4.5) suggests that this problem in fact extends even to non-lexical features, leading us to reject traditional within-corpus evaluation (i.e. crossvalidation). Here, we explore a novel approach to L1 identification which relies only on externally-derived lexical information. It involves deriving metrics from large weblog corpora for four L1s (Chinese, Japanese, Spanish, French), with the idea of lessening our reliance on scarce learner corpora. More specifically, we use the average ratios of (translated) word counts in different languages as indicators of interlanguage. If we see the unlikely English bigram *take coffee* in a learner text, our classification of that text will then depend on whether there are patterns of language in some L1 that could be the source of this L2 feature: among French, Spanish, Chinese, or Japanese, is there one language where we see a word that means *take* together with a word that

[6] Not to be confused with the idea of *interlingua* in machine translation.

[7] The work in this section is adapted from "Measuring interlanguage: Native language identification with L1-influence metrics" by Julian Brooke and Graeme Hirst, published in the *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (Brooke and Hirst, 2012b).

means *coffee*?

5.3.2 Method

The core of our method is the derivation of L1-transfer metrics. Given an L2 text, we derive an L1-transfer metric by averaging, across all relevant elements of a given type in the text, the ratio of the *potential prevalence* in contrasting L1 corpora (our training corpus). We will use the term *potential prevalence* to refer to counts that are filtered through some mapping; we cannot directly count L2 elements in L1 corpora, but we can count patterns that might produce them.

More formally, let $L_1, \ldots, L_p$ be the set of native languages we are interested in identifying, with corresponding corpora $C_1, \ldots, C_p$, and a small finite set of general feature types $T_1, \ldots, T_q$. As we will discuss in more detail later, our feature types include unigrams and bigrams. Our initial set of L1-ratios is then of size $p \times (p-1)$ i.e. one for each feature type for each pair of non-identical languages. For the moment, we assume a function P that provides a *potential prevalence value* for any given textual element e_{ij}, of type T_i, in some L1 corpus C_k, i.e. $P(e_{ij}, C_k) \rightarrow \mathbb{N}$. For a given element e_{ij}, we calculate its potential prevalence ratio $R_{lm}(e_{ij})$ for languages L_l, L_m as

$$R_{lm}(e_{ij}) = \log \frac{P(e_{ij}, C_l)}{P(e_{ij}, C_m)}$$

Note that the use of logarithms ensures that the two potential prevalence ratios derived from any languages are symmetric, $R_{lm}(e_{ij}) = -R_{ml}(e_{ij})$. Next, for all elements of type T_i in a given source text (the set E_i), we calculate the value of a feature $f_{lmi} \in F$ (corresponding to L_l, L_m, T_i) as the average of all the prevalence ratios for all relevant elements in the text:

$$f_{lmi}(E_i) = \frac{\sum_{e_{ij} \in E_i} R_{lm}(e_{ij})}{|E_i|}$$

Then, we define our set of L1-influence metrics V based on a combination of these basic features by language. A particular L1-influence metric v_{li}, l and i as above, is given by:

$$v_{li}(E_i) = \sum_{m=1}^{p} f_{lmi}(E_i)$$

Intuitively, each basic ratio in F provides an indication of whether a text is patterning more like one of two languages, while the set of L1-influence metrics V provides an indication of how much a text is patterning like a particular language in contrast with all other languages. Finally, we normalize these metrics in the context of the test corpus, so they all have the same standard deviation. For some text with textual elements E_i:

$$v'_{li}(E_i) = \frac{v_{li}(E_i) - \overline{v_{li}}}{\sigma_{v_{li}}}$$

A text is classified as the language L_c with the highest normalized influence metric, i.e.

$$c = \arg\max_{l} v'_{li}$$

The above provides an abstract basis for our classification using L1-influence metrics. However, we need to define the potential prevalence function, which depends directly on the type of feature T being extracted. Our main feature is what we call *boundary bigrams* (or just bigrams), which correspond to the L2 (translated) bigram associated with two consecutive words in an L1 corpus. Let us consider some L1 corpus C, with tokens $w_1 \ldots w_n$, each of which has some (possibly empty) set of translations $t_i = t_{i1}, \ldots, t_{ij}$, with each t consisting of one or more words in the target (L2) language, say $t_{ij1} \ldots t_{ijm}$. Then the potential prevalence function for the boundary bigram feature T_1 for an element e_{i1} corresponding to a ordered pair of target words $(t't'')$ is the count, across all adjacent words $w_i \ldots w_{i+1} \in C$ and across all their potential translations $t_{i1}, \ldots, t_{ij}, \ldots, t_{im}$, $t_{(i+1)1}, \ldots, t_{(i+1)k}, \ldots, t_{(i+1)l}$, of the number of instances where,

$t' = t_{ijq}$ and $t'' = t_{(i+1)k1}$, given $|t_{ij}| = q$. That is, a count of all the instances where the last word of one of the translations of some $w_i \in C$ is equal to the first word of the bigram, and the first word of one of the translations of $w_{i+1} \in C$ is equal to the second word of the bigram.[8] For instance, consider the French phrase *prends un café*, with a (partial) list of translations for each word as below:

w_i	*prends*	*un*	*café*
t_{i1}	take	a	coffee
t_{i2}	hold	an	java
t_{i3}	go by	one	cafe

An appearance of this phrase in a corpus would generate a boundary bigram count for *take-a*, *a-coffee*, *take-an*,..., *by-a*, *by-an*,..., *one-cafe*. They are boundary bigrams because we only consider the bigrams that straddle word boundaries (not *go-by*, for instance); assuming a reliable bilingual lexicon, within-word bigrams (when they occur) will involve only correct usage of the L2, but we intend boundary bigrams to find lexical patterns that reflect transfer from the L1.

A related way of using L1 corpora is to derive information via the use of *k-window collocational pairs*. These k-window collocational pairs differ from boundary bigrams in three key ways: first, they do not require strict adjacency, which is to say that for an integer k, w_i, $w_j \in C$ are considered k-window collocations if $|i - j| \leq k$. Second, we consider only those translations of length 1, i.e. only t_{ij} s.t. $|t_{ij}| = 1$. Third, collocational pairs are unordered, i.e. the sequence $w'w''$ will result in the same collocation counts as the sequence $w''w'$. Otherwise the potential prevalence function for collocational pairs is similar to boundary bigrams, a count of target language word pairs over all the words and all the translations of these words in the

[8] We also tested using pointwise mutual information as a potential prevalence indicator for boundary bigrams, but it was not as effective as raw counts. More generally, a probabilistic interpretation of potential prevalence assigns far too much probability mass to nonsense bigrams we will never see in actual texts.

L1 corpus. In our *prends un café* example, the 2-window collocational pairs include all combinations of all the single word translations, e.g. (*coffee*, *take*), (*hold*, *java*), but not anything with *by* or *go* since these are part of a multiword translation. Here, we only test 2-window collocational pairs.

For the unigram feature type, we simply count all target words (t_{ijk}) in all translations for all tokens in the corpus. For POS unigrams, bigrams, and trigrams, each word w_i is given a corresponding POS tag p_i, and we count sequences of these POS tags, and then map them to a single, coarse-grained tag set consisting of nouns, verbs, adjectives, adverbs, conjunctions, pre/postpositions, pronouns, numbers, punctuation, and the catch-all category of (other) function words, so that these counts can be compared across L1s. For combined unigram and bigram counts, we sum the potential prevalence ratios derived for each feature.

Not all elements of the text are equally useful for L1 classification. We posited that classification would be better if commonly occurring features of English were filtered, since these may vary randomly across L1 and produce noise. We implement this by fixing a maximum n-gram count, as derived from an independent corpus, for the elements used to calculate the L1-influence metrics. Appropriate thresholds were selected by optimizing in the held-out development set. We exclude proper nouns, which can of course be useful for L1-identification but should not be attributed to language transfer, which is our main interest here.

5.3.3 Data and resources

The data and resources used in this work can be divided into four categories: the (L1) corpora for deriving potential prevalence, resources for analysis of these corpora (e.g. segmenters, taggers), bilingual lexicons, and evaluation resources. For the L1 data, we choose to draw primarily from a single web corpus, the ICWSM Spinn3r dataset (Burton et al., 2009), which, although primarily an English corpus, also contains a large number of blog posts in other lan-

guages. There is a great deal of variation in the amount of data available for each language; for consistency, we choose a fixed length sample (100 million tokens, after segmentation) for each of the five languages. Chinese, however, was underrepresented with only 19 million tokens, and so we extracted additional blogs from a popular Chinese site.[9] In addition to ICWSM English data, we used the Google 1T 5-gram Corpus (Brants and Franz, 2006), which includes counts based on one trillion tokens from the web, for our count thresholds.[10]

For the European languages it was possible to use simple heuristics for tokenization, while for Chinese we needed a special segmenter: we employed the Stanford Chinese segmenter (Chang et al., 2008a), in the Chinese Treebank tagset mode. For Japanese, the MeCab morphological analyzer[11] served as our segmenter as well as part-of-speech tagger. For the other languages, POS tagging was carried out using the Tree Tagger (Schmid, 1995) and the associated parameter files for each language.

We did not have immediate access to sufficiently large machine-readable bilingual dictionaries for any of the (non-English) L1s, so we took advantage of the various websites which offer free online bilingual translations. Over the course of several months, we slowly and politely queried these websites for English translations of words that appeared often (at least 5 times)[12] in the corresponding subcorpus. For Chinese, we used *iciba.com*, for French *larousse.fr*, for Spanish *spanishdict.com*, and for Japanese *jisho.org*; our choice of websites was based on dictionary quality, ease of extraction, and, in particular for the European languages, the ability to deal with inflected forms, i.e. to find their corresponding lemma without need for additional lemmatization on our part. Although we attempted to keep the size of the dictionaries comparable, in terms of lemmas the Chinese and Japanese lexicons are markedly larger than the

[9] http://www.sina.com

[10] We summed relevant trigram counts to get our thresholds for the 2-window collocations.

[11] http://mecab.sourceforge.net/

[12] All of our query-derived lexicons in fact may have more than just those words appearing 5 times in the corpus, but this is the last cutoff point that all dictionaries reached. We do not, however, believe there is much benefit to be gained from further extraction, since such rare words rarely have definitions in the online dictionaries.

French and Spanish ones;[13] if inflected forms are considered, however, the European-language lexicons are larger.

For all languages, we ignored translations longer than three English words, as we found that many of these were explanations rather than translations. Some very common words in some lexicons had only explanatory entries; for these (fewer than 10 in each lexicon) we manually inserted a direct translation based on examples or, in the case of certain particles, left them with an empty translation. The translations of verbs and nouns were generally in base form, which would have resulted in only uninflected bigrams; instead, we used the part-of-speech tagging to create simple correspondences between forms in the L1 and inflected forms in English. For instance, plurals in French are translated into plural forms in English; for Chinese, however, which does not mark number on most nouns, both English forms are included as potential translations. All of the dictionaries categorized their translations by part of speech, and in general we used the translations for only the part of speech as given by the tagger, though all translations were used if that strategy failed.

Our first evaluation corpus is the International Corpus of Learner English (ICLE), version 2 (Granger et al., 2009); for each of the 4 languages investigated here, we used the first 50 texts in each subcorpus for development, and the next 200 for testing. Our second evaluation corpus is our new Lang-8 corpus (Brooke and Hirst, 2011); since the average entry in the Lang-8 is significantly shorter than those in the ICLE (about 150 tokens), we concatenate multiple entries together to form our 'texts' of roughly the same length as those in the ICLE, also 200 for each language.[14] Our third corpus is the First Certificate in English (FCE) corpus (Yannakoudakis et al., 2011). The set we use here consists of only 50 texts per language, and the average length of the texts is roughly half of the other two corpora; thus we expect classification to be harder.

[13] Chinese: 109,061, Japanese: 85,867, Spanish: 26,627, French: 26,495.

[14] Although this may appear to be a fairly small sample of this corpus, in fact the 200 texts nearly exhaust the data available for the two European languages.

Table 5.2: Native language classification results

Configuration	Accuracy (%)					
	ICLE texts		Lang-8 texts		FCE texts	
	No Filter	**w/Filter**	**No Filter**	**w/Filter**	**No Filter**	**w/Filter**
Guessing baseline	25.0	25.0	25.0	25.0	25.0	25.0
Unigrams	43.5	44.6	26.0	26.9	22.0	22.5
Bigrams	42.9	**48.3**	**36.4**	**39.2**	28.5	29.0
2-window collocations	32.1	46.9	31.9	38.3	29.5	**32.0**
POS unigrams	25.0	30.1	26.4	30.0	**32.5**	26.0
POS bigrams	17.0	25.3	26.9	29.1	25.5	27.0
POS trigrams	16.5	28.8	27.8	28.4	23.0	26.5
Unigram + Bigrams	**44.2**	46.2	27.9	30.1	24.3	23.7

5.3.4 Evaluation

Table 5.2 contains the L1 classification results for the various feature types and evaluation corpora. The boundary bigram and *k*-window collocations are obviously the most useful feature types; their performance is consistently well above chance, even without filtering. By comparison, the POS features do not appear to transfer properly and perform often near or even below chance, perhaps because the sequences in which POSs appear are just simply too language dependent. The effectiveness of unigram features vary widely: in the ICLE, they are roughly as good as bigrams, but in the FCE they are worse than guessing. We suspect that these variations may reflect a fundamental difference in the nature of the two corpora: the short answers in the FCE are constrained to a very restricted topic and genre—letters expressing gratitude at winning a prize—which may limit the extent to which vocabulary choice can distinguish among L1s. It is therefore the choice of which words are put together that is particularly telling, reflecting transfer from the L1.

One very clear result is the effect of filtering: in nearly every case, filtering out elements that were common in English improved classification accuracy. This effect is most pronounced in the ICLE (from which we also took our development set), but it is visible in the other two

Table 5.3: Confusion matrix for best ICLE result

Native	**Classified as**			
Language	Chinese	Japanese	French	Spanish
Chinese	103	42	27	28
Japanese	41	111	18	30
French	15	30	86	69
Spanish	13	33	68	86

corpora as well. The bigram threshold was 10^6 appearances in a corpus with roughly 10^{12} bigram tokens. One negative result is that the features do not appear to combine well; in general, summing unigram and bigram metrics did not improve performance.

Table 5.3 contains the confusion matrix for the bigram L1-influence metric in the ICLE, our best result. The Asian languages are the easiest to distinguish, while the two closely related European languages are distinct from the Asian but often misclassified as each other. This is exactly what we should expect given our knowledge about how the languages are related to each other. We suspect performance would be much higher if we had not included languages that are so closely related to each other as well as English (that is, French and Spanish), though even these two languages are distinguished better then chance.

We also looked at the individual bigrams that contributed to the metrics, in particular those with very high or low potential prevalence ratios. Among the most telling features for Chinese, we noticed a number of Chinese-influenced adjective-noun collocations (e.g. *main income*, *medium industry*), but there were also syntactic errors of number (e.g. *they depends*). The patterns were less clear for European languages like French, though we noted certain verb-preposition combinations (e.g. *tolerated to*, *witnessing in*) that seemed to be cases of language transfer. There was also a great deal of noise, which might be eliminated by further filtering, for instance focusing only on specific POS patterns.

5.3.5 Discussion

In this section, we have presented a method for using native language corpora as a source of information for native language identification in non-native texts. In particular, our approach relies on the phenomenon of language transfer, where patterns of the L1 intrude into the L2. The results offered here are well above chance, though they are not good enough for us to conclude that this method alone is sufficient; notably, this method would not identify a general lack of lexical diversity in a text, and our handling of syntax is fairly crude. However, there are aspects of our method that make it distinct from traditional machine-learning approaches: in particular, our metric can provide a small set of features that may represent a huge number of rare (but telling) events that might otherwise be filtered out by feature selection. Our method also offers an explicit connection between L2 forms and the L1 forms that created them; this information could be used to improve automated error correction.

Our method relies on an averaging of ratios of translation-based counts for the different L1s across appearances of relevant *n*-grams in the text. Before we move on to supervised methods in the next section, we stop to justify why we took this approach, and not one that involved using the L1 information to create a prototype distribution (i.e. a normalized vector of *n*-gram counts) for each L1, that could be compared to a similar distribution for L2 texts. We have already mentioned (in footnote 8), why we are hesitant to assign any probabilistic interpretation to the counts from our L1-transfer method: there are a huge number of counts (probably the vast majority) which are simply garbage, the result of indiscriminate translation of all possibilities; the size of this effect in each L1 is dependent on irrelevant factors such as the thoroughness of each bilingual lexicon. We could perhaps side-step this problem by creating an L1-distribution based only on *n*-gram features that actually appear in the corpus (throwing the other counts away), though this is a bit strange. Creating an *n*-gram distribution of each L2 text is also potentially problematic: Although the appearance of a given lexical feature might

be important, the number of times that feature is used is more likely to be a reflection of topic or the requirements of syntax than L1 influence, and there are of course many very relevant features that are also missing simply because the author had no opportunity to use them; the random effect of these absences could easily overwhelm the resulting output. This in turn might be countered by use of a metric like KL-divergence (of the L1 corpus distribution from the L2 text distribution) which essentially disregards low-probability elements, but which would also give high weight to commonly appearing elements in the L2 texts, which, based on our filtering results, does not seem to be the right approach here. Overall, although we cannot entirely rule out that such an approach might yield dividends, particularly in combination with additional supervision, there are important ways in which directly comparing such distributions in an unsupervised way is rather unnatural. We believe that the ratios for individual *n*-grams across L1s do directly reflect the presence or absence of L1 transfer, though admittedly there might be more principled ways to derive these counts and better ways to combine the resulting ratios, ones that might remove influences other than language transfer (in particular, random noise). A more important question, though, is how this information of this sort might be integrated successfully with statistical systems of the kind in discussed the next section.

5.4 Cross-corpus supervised classification using lexical *n*-grams

5.4.1 Introduction

Though a wide range of feature types has been explored for NLI—with conflicting results—the evaluation of these feature sets has been fairly uniform: training and testing in one of several small corpora of learner essays (Granger et al., 2009; Yannakoudakis et al., 2011; Lüdeling et al., 2008), which are unfortunately quite expensive to collect.[15] A notable problem with

[15]The work in this section is based on "Robust, lexicalized native language identification" by Julian Brooke and Graeme Hirst, published in the *Proceedings of the 24th International Conference on Computational Linguistics* (Brooke and Hirst, 2012d).

these corpora with respect to native language identification, however, is a clear interaction between native language and essay topic. Generally speaking, the solution in previous work has been to avoid the use of lexical features that might carry topical information, limiting feature sets to syntactic and phonological phenomena. There are two reasons to be critical of this approach. First, there are almost certainly kinds of language transfer (Odlin, 1989), i.e. transfer related to lexical choice, that are being overlooked. Second, and more troubling, is that avoiding the lexicon is not fully effective as a means of countering the effects of topic: some recent work indicates that variation in topic also has significant influence on non-lexical features (Golcher and Reznicek, 2011), calling into question the reliability of previous results that assume otherwise.

The approach we present here resolves this tension by requiring training and test sets that are independently sampled. Although corpora may have some form of confounding variation that may artificially inflate or (in some cases) lower performance relative to other samples from the same corpus, any variation that is consistent across very distinct corpora is likely to be a true indicator of L1. Although we test on the typical essay corpora used by other researchers, we train on the Lang-8 (see Section 5.2), a large but messy corpus of journal entries scraped from a language learner website. Without the distraction of (irrelevant) topic bias, we can test the efficacy of lexical features, including *n*-grams and dependencies. We also test a number of options at the level of the classifier, most notably a multiclass support vector machine (SVM) decision-tree classifier that leverages the genetic relationships among languages, and a simple but elegant method for adapting an SVM classifier to the test corpus without integrating the confounding variation found there. Our best classifier with lexical and syntactic features provides results that compare well with previously-reported single-corpus performance; we also present, however, evidence that calls into question the validity of these previous results, showing that topic bias within the corpus is having a major effect and that

Table 5.4: Number of texts in learner corpora, by L1.

L1	Corpus		
	Lang-8	**ICLE**	**FCE**
Japanese	59156	366	81
Chinese	38044	982	66
French	1414	347	146
Spanish	3080	251	200
Italian	1072	392	76
Polish	1549	365	76
Russian	7159	276	83

indeed the performance of models built in the topic-biased ICLE corpus is not robust, regardless of the features chosen.

5.4.2 Corpora

Our training corpus is the Lang-8 corpus, discussed in some detail earlier (Section 5.2) The average length of an entry is about 150 word tokens after our pre-processing; since these texts are relatively short compared to our test sets, for our purposes here we append consecutive short texts of writers with the same L1 (often the same author) until they are at least 250 tokens in length, which results in an average length of 431 tokens.

Our main test corpus is the International Corpus of Learner English (Granger et al., 2009). As already mentioned, a major problem with the ICLE is topic variation, which is both unnaturally strong and often arbitrary. The average text length in the ICLE is 617 words. Our second test corpus is the FCE. The average length of the texts in the FCE corpus is 428 words, or about 200 words less than the ICLE.

For this study, we selected the seven languages which had sufficient numbers in all three corpora, i.e. at least 1000 texts in the Lang-8 corpus, 200 texts in the ICLE, and 50 texts in the FCE. Table 5.4 shows, for each L1, the number of texts present in each corpus. For testing in the ICLE, we use 200 from each set, and a separate set of 50 per L1 is used for our bias

adaptation method. For testing in the FCE, we use 50 texts, and 15 texts for bias adaptation.

5.4.3 Classifier experiments

We split our main experiments into two parts. In our initial investigation, we found that using the full set of feature types, to be described later, provided near-optimal results. Given that exploring the exhaustive set of combinations is not feasible in this space, we elect to first take the full feature set as fixed and turn our attention to higher-level classifier options, establishing the best among those options before we proceed with a feature analysis.

Our experiments included testing the following options:

Balanced training (bal) vs. cost weight (cost) Statistical classifiers generally depend on having similar class distributions in training and testing sets, an assumption which is violated here. There are two simple ways to handle this problem: either balancing the training sets by discarding extra training data, or training the classifier with using different cost weights for different classes, promoting classification of rarer classes to the level expected in the (balanced) test data. We use the cost weight equation from Morik et al. (1999).

Binary (bin) vs. frequency (freq) features Previous work has mostly used normalized frequency rather than binary occurrence in a text as the feature value used for classification; Wong and Dras (2011) are an exception, but they do not justify that choice.

SVM vs. MaxEnt classifier Support vector machines were a popular option in previous work, but Wong and Dras (2011) report better performance with a Maximum Entropy (MaxEnt) classifier. A full discussion of these two machine learning methods is omitted here, though we note that (pairwise) SVMs are generally conceptualized as a hyperplane which maximizes the margin between classes in the feature space, while MaxEnt is a multinomial logistic model

built by constrained maximization of the probability of the training data. For SVM classification (see below), we use LIBLINEAR (Fan et al., 2008), which is optimized for linear kernel classification of large datasets; except as explicitly mentioned below, we present results using default parameter settings (which were found to give good results). Feature vectors are normalized to the unit circle (Graf and Borer, 2001). For MaxEnt we follow Wong and Dras (2011) in using MegaM.[16]

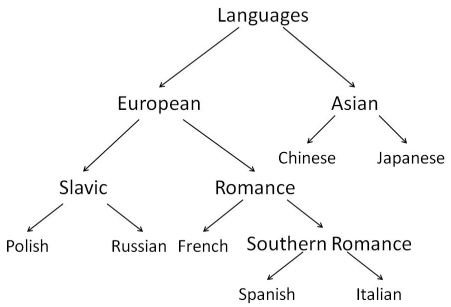
Regularization parameters In the context of building a robust classifier for cross-corpus classification, the regularization of the model (Alpaydin, 2010), i.e. the degree to which the classifier increases in complexity to fit the training data, is of obvious relevance. For SVMs, the key parameter is C, which controls the penalty for misclassified examples in the training set: a large value of C means these errors have a higher influence on the objective function, promoting more complex models that minimize error but possibly results in overfitting. For the MaxEnt classifier, the λ parameter controls the influence of a Gaussian prior on the feature weights: low values of λ correspond to an imprecise prior, allowing the feature weights to fit the data. We tuned the corresponding parameter for each classifier configuration using 7-class task performance in the development set for each test corpus.[17]

Multiclass SVM type While MaxEnt has a natural multiclass interpretation, an SVM decision plane is appropriate only for binary choice. A standard approach to multiclass SVM is to combine multiple pairwise SVM classifiers (Hsu and Lin, 2002). Two general options in this vein are *one vs. one* (1v1), where $n(n-1)/2$ individual classifiers (for n classes), each trained on one pair of classes, are combined, and *one vs. all* (1va), where n classifiers are trained by separating one class from all the others. The winner of 1va is obviously the class with the

[16] http://www.cs.utah.edu/~hal/megam/

[17] Since the C parameter is selected once for each configuration based on the 7-class task, some results that we would otherwise expect to be equivalent, e.g. the 2-class SVM classifiers, actually vary slightly.

Figure 5.1: Binary decision tree for SVM experiments



highest margin (distance from the decision plane), but for 1v1 it is typically the class which is chosen by the most classifiers (ties are broken in favor of the highest margins). A third, novel option is made possible by the genetic relationships among languages in our test set: an SVM binary decision tree (tree), presented in Figure 5.1.[18] Note that tree classifiers have a significant performance advantage over both 1va and 1v1 classifiers with respect to the number of classifiers required ($n-1$), and an advantage over the 1va classifiers with respect to the average size of the training sets used to build those classifiers. Finally, Crammer and Singer (2002) have proposed a multiclass SVM classifier based on class prototypes (pro) rather than hyperplane boundaries, and we also test this option (as implemented in LIBLINEAR).

Bias adaptation, pairwise (adS) Since there are significant differences in the genre, domain, and quality of texts across our training and test corpora, some form of domain adaptation (Daumé and Marcu, 2006; Bruzzone and Marconcini, 2010) would almost certainly be helpful. However, even unsupervised forms of transfer learning (Pan and Yang, 2010) are likely to take advantage of those confounding factors that prompted us to reject within-corpus evaluation; we believe that any change to the feature weights based on samples from the same corpus that our test set is drawn from is ultimately self-defeating in this context. However, there is one

[18]There is some controversy in the literature about the genetic relationship amongst Romance languages; see the discussion by Kochmar (2011).

key parameter to these that is not a feature weight: the bias. In pairwise SVM, changing the bias slides the hyperplane, changing only the total number of positive (or negative) features required to make a classification, not the individual influence of a particular feature (i.e. the sign of a feature weight). With respect to its effect (changing the balance of classes), it is closely related to our cost factor option above; however, whereas the cost factor is a parameter used during training, we shift the bias using our own iterative process after the model is built, using a sample from the same corpus as the test set (a development set).[19] Our algorithm is as follows: we first initialize our step size to the absolute value of the original bias, and then we iteratively modify the bias, adding or subtracting the present step size such that we are moving in the direction of a distribution where the ratio of classes predicted in our development set is the same as in the final test set, reclassifying the data after each step.[20] If we overshoot the desired ratio, we halve the step size, and continue until we reach the desired ratio or the predicted ratio does not change for 10 iterations. We do this separately for each test set with the corresponding development set.

Bias adaptation, multi (adM) The MaxEnt and SVM prototype classifiers also have bias terms that can be optimized, but unlike the pairwise classifiers they cannot be dealt with one at a time; optimizing the bias for one class will affect the others in unpredictable ways. We proceed with the same basic algorithm as the pairwise classifier, but we do this for all bias terms simultaneously, i.e. all biases are adjusted in a single step. Each bias has a separate step size, and the optimization ends when the entire distribution is correct or nothing has changed in 10 iterations. We also implemented this for SVM 1va, i.e. interpreting it as a single multiclass

[19] Admittedly, we could accomplish this with additional parameter tuning, but there are both practical and principled reasons for doing it this way: it is much faster to modify the biases directly rather than retraining the model, and, more importantly, we want to preserve the original feature weights; we require that they do not reflect exposure to the confounds of the testing corpus in any way.

[20] This requires knowledge of that distribution. However, it is otherwise unsupervised in that we are only concerned with the distribution of predictions: we do not use the true class values except to create the appropriate subsets for the SVM 1v1 and SVM tree classifiers.

Table 5.5: Native language classification accuracy (%) for varying classifier options. Bold indicates best result in column, italics indicates difference from the pivot classifier (11).

Configuration	Asian		European		All	
	ICLE	FCE	ICLE	FCE	ICLE	FCE
Chance baseline	50.0	50.0	20.0	20.0	14.3	14.3
(1) SVM *1v1* cost bin	95.2	86.0	50.0	40.4	58.7	50.3
(2) SVM *tree* cost bin	95.2	86.0	48.7	41.3	59.4	49.4
(3) SVM 1va cost bin	**96.5**	86.0	54.8	44.0	61.6	50.8
(4) SVM *pro* cost bin	95.0	85.0	55.6	42.8	62.4	50.8
(5) *MaxEnt* cost bin	95.0	85.0	56.6	44.8	63.7	42.3
(6) SVM *tree*-adS cost bin	95.2	**88.0**	64.4	57.2	73.7	57.4
(7) *MaxEnt*-adM cost bin	95.0	86.0	68.2	64.4	74.0	60.8
(8) SVM *1v1*-adS cost bin	95.5	88.0	67.9	66.8	74.2	65.7
(9) SVM 1va-*adS* cost bin	95.0	88.0	71.6	67.6	77.8	**66.5**
(10) SVM *pro*-adM cost bin	95.7	87.0	71.1	66.4	77.3	64.0
(11) SVM 1va-adM cost bin	95.0	86.0	**71.7**	**68.0**	**78.0**	65.7
(12) SVM 1va-adM *bal* bin	79.8	75.0	63.1	60.4	66.8	59.1
(13) SVM 1va-adM cost *freq*	95.2	83.0	66.8	57.6	74.9	53.1

classifier rather than a set of pairwise classifiers.

Table 5.5 shows the results of our experiments. In addition to the full 7-language task accuracy (the 'All' columns), we also present results classifying the two major subgroups; note that these are distinct tasks, e.g. for European it is the accuracy of a 5-language task, not the accuracy of the classification of those 5 languages within the 7-language task (see Figure 5.1 for our language classification schema). However, in our discussion, we focus on results for the full 7-language task. The upper part of Table 5.5 includes various key classifier options, ordered by their 7-way ICLE accuracy, while the bottom includes other options; the best classifier (11) is used as a pivot between the two.[21] The aspect(s) of the configuration that are different from the pivot are in italics, and the best results in each column are in bold. For each classifier, we report the results using the best C or λ values from an initial series of runs using the development set.

Unsurprisingly, we see better results when we use all the data at our disposal (11), rather

[21]The effects of the options in each of the two parts of the table are fairly independent, so for simplicity of presentation we test them separately.

than forcing balanced test cases (12). This result is useful, though, because it indicates that our consistently high performance in distinguishing Chinese and Japanese elsewhere in Table 5.5 is a result of that extra data, and not other factors, i.e. the fact that unlike our other language groupings, Chinese and Japanese do not belong to a single genetic language family (Comrie, 1987). Also clear is the preference for binary (11) rather than frequency-based (13) feature values: one possible explanation is that, in these relatively short texts, there is high variability in normalized frequencies, and a simpler metric, by having less variability, is easier for the classifier to leverage. In general, slightly less regularization (high C, low λ) values were preferred, though most were reasonably close to the default values; tuning made little difference, particularly for the SVM classifiers.

Between the two main classifier types, the MaxEnt classifier was, with the appropriate choice of λ (5), the best performing classifier in the ICLE when no bias adaptation was used; it was, however, worse than almost all of our SVM options in the main 7-language classification task when bias tuning was allowed (7). This does not appear to be a failure of the adaptation algorithm, but rather a real distinction between the two classifiers: our experience is that the SVM classifiers are less robust, i.e. more prone to errors when training and test sets differ significantly, but they can be easily recalibrated for optimal performance with a relatively small amount of information. Here, we show that changing the bias alone is enough for major gains across all the SVM types (6,8–11), results which are statistically significant.

Our novel binary tree classifier (2,6) is competitive but ultimately performs poorly compared than other options, suggesting that the simplicity of the classifier does come with a trade-off in performance. The 1va classifiers (3,9,11) are consistently better than 1v1 (1,8), while the performance of the prototype-based SVM (4,10) is nearly indistinguishable from 1va. This is somewhat surprising, since we might expect a 1v1 or prototype approach to be able to better deal with the commonalities and differences among languages than the 1va, which

lumps diverse languages into a single 'other' category. With respect to the 1va classifier, it does not seem to matter much whether pairwise (9) or single classifier (11) bias tuning is used; the latter gave us the best 7-class performance in the ICLE (and we use it as our best classifier), but the former gave slightly better performance in the FCE. In the ICLE, the difference between the best bias-adapted 1va classifier and the 1v1, tree, and MaxEnt classifiers is statistically significant (χ^2 test, $p < 0.001$).

5.4.4 Feature analysis

Our model includes the following feature types:

Function words A common feature in stylistic analysis. Our list of 416 common English words comes from the LIWC (Pennebaker et al., 2001).

Character *n*-grams (unigrams, bigrams, and trigrams) For bigrams and trigrams, the beginning and end of a word are treated as special characters.

Word *n*-grams (unigrams and bigrams) Note that word *n*-grams are a superset of function words. Punctuation is included.

POS *n*-grams (unigrams, bigrams, and trigrams) POS tagging is provided by the Stanford Parser V1.6.9 (Klein and Manning, 2003), also used by Wong and Dras (2011).

POS/function mixture *n*-grams (bigrams and trigrams) Wong et al. (2012) report better results with POS *n*-grams that retain the identity of individual function words rather than using their part of speech.

CFG productions Context-free grammar production rules, as provided by the Stanford parser. Lexical production rules are not included.

Dependencies Dependencies consist of two lexical items and the syntactic relationship between them. Also produced by the Stanford parser (de Marneffe et al., 2006).

Syntactic Features POS *n*-grams, POS/function mixture *n*-grams, and CFG productions.

Lexical Features Word *n*-grams and dependencies.

Proper Nouns Not actually a separate feature, proper nouns are included by default in character and word *n*-grams as well as dependencies. They are obviously relevant to the task, but there are applications (e.g. forensic profiling) where they might not be appropriate, since they do not directly indicate language transfer from the L1 but rather reflect real-world correlations between native language and country of residence, etc. Here, we report results with all proper nouns excluded from consideration for all relevant features.

Feature Selection Wong and Dras (2011) tested feature selection based on information gain, but it provided no improvement in performance. For practical reasons, we have included by default a simple frequency-based feature selection; only features that appear in 5 different texts in the training set are included. Even with this restriction, our feature set has almost 800,000 features. Here, we test the effect of a higher frequency cutoff (at 20), and limiting our set to features with positive information gain.

Again, we focus on the results of the full 7-language task (the 'All' columns). Clearly, all the feature types can be used to distinguish native language: each of the results in Table 5.6 is well above a chance baseline, though function words (1) and character *n*-grams (2) give a fairly modest performance individually. Compared to these, production (5) rules are markedly more useful, a result which is compatible with the conclusions of Wong and Dras (2011). Nonetheless POS (3) and in particular mixed POS/function words *n*-grams (4) offer even better

Table 5.6: Native language classification accuracy (%), by feature set. Bold indicates best result in column.

Features	Asian		European		All	
	ICLE	FCE	ICLE	FCE	ICLE	FCE
Chance baseline	50.0	50.0	20.0	20.0	14.3	14.3
(1) Function words	72.7	71.0	40.3	37.2	35.6	36.0
(2) Character *n*-grams	78.3	63.0	37.5	28.8	37.4	22.6
(3) POS *n*-grams	86.8	78.0	47.9	50.0	52.9	44.3
(4) POS/function *n*-grams	93.3	85.0	60.6	56.8	67.4	59.4
(5) CFG productions	78.5	72.0	46.9	43.2	49.7	41.1
(6) Dependencies	94.0	79.0	49.8	46.8	61.4	45.1
(7) Word *n*-grams	94.3	89.0	71.1	66.8	77.1	68.3
(8) Syntactic Features	94.3	87.0	60.1	61.2	68.1	65.1
(9) Lexical Features	95.2	86.0	71.0	67.6	77.8	67.1
(10) Lexical+Syntactic	**96.0**	90.0	72.3	66.4	78.4	68.2
(11) All features	95.0	86.0	71.7	68.0	78.0	65.7
(12) (4)+(7)	95.5	90.0	**72.5**	66.8	**79.3**	**70.0**
(13) (4)+(7), no proper nouns	94.5	87.0	69.6	67.2	76.5	65.7
(14) (4)+(7), df ≥ 20	95.0	86.0	71.3	**68.4**	77.3	65.4
(15) (4)+(7), IG > 0	89.5	**93.0**	69.5	66.4	76.5	65.7

performance, despite being somewhat simpler. Compared to the latter of these, the usefulness of lexical dependencies (6) is muted, and shows a very inconsistent performance across the two test sets. Word *n*-grams (7), however, alone account for almost all of the accuracy we see when all features are combined.

Adding the POS features and CFG productions (8) generally boosts performance, suggesting that the syntactic features may not be entirely redundant, while the combination of the lexical features also provides a small improvement in the 7-language ICLE task, though the FCE is worse (9). Further adding the syntactic features to the lexical features increases performance for most of the tasks (10), while including character *n*-grams tends to degrade performance (11). Finally, we exhaustively tested feature combinations and found that the best performing for the 7-language task used only the two best individual feature types, POS/function word mixtures and lexical *n*-grams, though the differences among all the options containing lexical

n-grams are not statistically significant (12).

When we remove proper nouns (13), there is a modest drop in performance, indicating that they had some positive role in the classification, but the benefits of using lexical features goes well beyond proper nouns. Additional frequency-based feature selection (14) has a small, mostly negative effect, as does restricting features to those with positive information gain (15). In general, we see no evidence that a simpler model is preferred in this case, though if speed is a concern one can be used without too much loss.

We also looked briefly at the individual lexical features that were useful based on their information gain in the training set. One thing that was immediately evident is that some common, entirely correct English words and expressions were extremely helpful for distinguishing native languages. For example, the phrase *decide to* was ranked high: we note that in at least one language in our set (French), a closely analogous cognate construction *decider de* exists, whereas another language, Chinese, has no analogous construction, since the verb that most closely means *decide to* (*jueding*) is phonetically dissimilar, has no element corresponding to *to*, is more common as a noun, and in fact is pragmatically associated only with major decisions, often in a legal context (closer to the English *make a decision to*). By default, learners will prefer forms that correspond to those from their L1 (Odlin, 1989), and lexical features are key to identifying this kind of language transfer.

5.4.5 ICLE-training experiments

One of the primary motivations for our cross-corpus approach to NLI is the confounding variation found in the ICLE corpus. In this section, we turn to using the ICLE as a training corpus in order to highlight these problems, particularly those relevant to 'stylistic' features, which have been thought of as immune to these effects. The first experiment, the results of which are presented in Table 5.7, consists of two types of 2-fold cross-validation in the ICLE corpus: the

Table 5.7: ICLE within-corpus experiment classification accuracy (%), by feature set.

Features	Random	Segregated	Difference
Chance baseline	14.3	14.3	–
(1) Function words	58.0	46.7	−11.3
(2) Character *n*-grams	51.2	48.2	−3.0
(3) POS *n*-grams	83.3	72.2	−11.1
(4) POS/function *n*-grams	87.6	79.2	−10.4
(5) CFG productions	86.1	79.7	−6.4
(6) Dependencies	89.1	77.1	−12.0
(7) Word *n*-grams	94.3	81.3	−13.0
(8) All (1–7)	90.4	81.6	−8.8

first is standard, randomized cross-validation, while in the second, the two folds (of 700 texts each) are segregated by essay prompt; essays based on a given prompt are in one fold or the other.[22] For this we use the 1va classifier without any bias adaptation, which is unnecessary in the case of cross-validation.

Within the ICLE, we see in the 'Difference' column of Table 5.7 the consistent effects of essay prompt on classification, across all kinds of features. The effects on lexical features (6,7) are, not surprisingly, most pronounced, but other popular features are also implicated to varying degrees. The effectiveness of various features under both conditions roughly mirrors the results in the previous section, though there are a few notable exceptions: for instance, production rules (5) were more useful here than in the Lang-8 trained cross-corpus experiments; this is interesting since many of the most recent results in the ICLE (Wong and Dras, 2011; Swanson and Charniak, 2012) make use of these grammatical features. Surprisingly, character *n*-grams were the least affected, a contrast from our preliminary work on ICLE topic bias (Brooke and Hirst, 2011), though there remains little doubt that they are inferior features for this task. Lexical *n*-grams are ultimately the most preferred feature (7), even when topic effects are partially[23] controlled for.

[22]This experiment is possible only in the ICLE, since titles in the Lang-8 are freely chosen by each writer, and there is little variety of prompts in the FCE.

[23]There are more pervasive topic and genre effects that segregating by prompt does not resolve. For instance,

Table 5.8: ICLE-training cross-corpus classification accuracy (%), by feature set.

Features	Lang-8	FCE
Chance baseline	14.3	14.3
(1) Function words	27.6	20.0
(2) Character *n*-grams	29.7	24.0
(3) POS *n*-grams	37.0	32.8
(4) POS/function *n*-grams	40.2	33.4
(5) CFG productions	32.5	31.4
(6) Dependencies	30.7	25.1
(7) Word *n*-grams	50.8	35.7
(8) All (1–7)	46.8	39.1
(9) Adaptor grammar *n*-grams	40.9	30.8

We also present cross-corpus experiments with the FCE and a language-balanced 150-text portion of the Lang-8 corpus as test sets. As with our training set, this test set consists of combined texts, this time with a minimum length of 500, making the texts of comparable length to those in the ICLE. We create another set of 50 texts for bias adaptation. In the latter experiment, we also include a special set of features: the POS/function mixture 5-grams which were selected by the adaptor grammars of Wong et al. (2012), providing superior performance over exhaustive enumerations. Since these features were derived from the ICLE, they could not be defensibly used in other experiments (i.e. with the ICLE as a test set), but we can test their usefulness here. Since the original experiment involved cross-validation, there are in fact 5 different sets; our set consists of the union of these sets.[24]

The cross-corpus results in Table 5.8 are strikingly lower than the within-ICLE results. They also compare poorly to our earlier cross-corpus results. Part of this difference is, of course, the effect of the much-larger Lang-8 dataset, though the balanced result in Table 5.5 (12), uses a very similar amount of data (as measured in tokens) from the Lang-8 but attains

a large number of the Japanese texts are personal narratives, each with a different title, while in the Russian texts there is a particular focus on the literature of various authors, and in the Chinese texts there is a discussion of the advantages or disadvantages associated with certain government policies.

[24]We originally intended to take the intersection, but in fact the intersection of the feature sets is empty; no single feature was useful in every fold.

a much better FCE classification accuracy (roughly 20% better). The POS/function mixture features (9) derived using adaptor grammars do reasonably well, but are only marginally better than exhaustive mixture features (4) in the Lang-8 test set, and are markedly worse than a number of other features in the FCE. Again, lexical *n*-grams (7) are obviously the best individual feature type.

5.4.6 Discussion

The results in the previous section highlight the problematic nature of within-corpus evaluation in general, and the inadequacy of the ICLE as a training corpus in particular. It is unclear to what extent previous results on this task are influenced by these effects, but we believe there is at least reason to be skeptical of some of the conclusions. In particular, sophisticated feature selection techniques which have been the focus of recent work may result in models which perform better in the ICLE, but which have little or no benefit beyond that particular corpus. We believe more attention should be paid to the overall validity of NLI experiments, rather than to specific technical approaches. One interesting open question is whether features such as proper nouns, which are of obvious but somewhat trivial benefit, should be excluded. Certainly, we would argue that lexical features in general are far too important to the task to simply be discarded; our experiments here suggest that their usefulness goes well beyond proper nouns and is not simply a reflection of topic.

Though higher performance is clearly possible using cross-validation, our Lang-8 trained model does reasonably well in both our testing corpora; the results are fairly consistent, and the difference can be attributed to the smaller size of the FCE texts. It is clear that factors such as the choice of classifier and the size of the dataset play some role, though the most obvious improvement came from the use of our bias adaptation technique, which uses a small amount of data from a test corpus to improve the model; this was particularly effective for

SVMs. Importantly, this method keeps the feature weights constant, a necessary precondition when the testing corpus has known arbitrary biases. Given the variation in text size, genres, and learner proficiency, some kind of adaptation is clearly necessary to get competitive results, though our experiments using it with the ICLE as training data suggest the method cannot overcome a problematic training set.

We note that our still sizable error rate on this task may in fact be due to a learner proficiency effect; on inspection, I found that some of the European texts were nearly indistinguishable from native writing. As suggested by the statistics provided in the ICLE manual (Granger et al., 2009), many of these learners are highly proficient, and thus they might have completely integrated the norms of their L2, making them legitimately indistinguishable. We also tested the correlation between essay scores in the FCE and our classification accuracy, and found a small negative correlation, suggesting that those who scored better were harder to classify; text length, though, was a confounding factor, since longer texts got better scores and are also easier to classify. Finally, we also noticed that French was the most consistently misclassified language, by a significant margin; this could be due, in part, to the historical connection between French and English that makes French L1 transfer somewhat less distinct, whereas distant languages like Chinese and Japanese are easy discerned, an effect we saw even when the training sets were balanced. In general, we think the relationship between proficiency, distances between languages, and L1 classification merits further study.

One important strength of the current work is the training dataset, which, unlike many learner corpora resources, is fully accessible via the web (and growing!). The coverage of European languages is poor, however, and since large amounts of data are necessary to fully leverage the potential of lexical features, one future direction would be to look for even more inexpensive ways of finding learner texts, perhaps by collecting English texts that appear on otherwise non-English websites. Armed with larger datasets, we would like to move beyond

classification of a handful of L1s, moving towards a system that can identify influence from a full range of common L1 backgrounds.

5.5 Using other corpora in the 2013 NLI shared task

5.5.1 Introduction

Our participation in the 2013 NLI shared task (Tetreault et al., 2013) follows on directly from our work exploring cross-corpus evaluation, an approach that is now becoming fairly standard alternative in relevant work (Bykh and Meurers, 2012; Tetreault et al., 2012; Swanson and Charniak, 2013).[25] As discussed earlier, our promotion of cross-corpus evaluation in NLI was partially motivated by serious issues with the most popular corpus for native language identification work up to now, the International Corpus of Learner English (Granger et al., 2009). The new TOEFL-11 (Blanchard et al., 2013) used for this NLI shared task addresses some of the problems with the ICLE (most glaringly, the fact that some topics in the ICLE appeared only in some L1 backgrounds), but, from the perspective of topic, proficiency, and particularly genre, it is necessarily limited in scope (perhaps even more so than the ICLE); in short, it addresses only a small portion of the space of learner texts. Our interest, then, continues to be in robust models for NLI that are not restricted to utility in a particular corpus, and in our participation in this task we have focused our efforts on the open-training tasks which allow the use of corpora beyond the TOEFL-11.

5.5.2 Basic model

In our recent work on cross-corpus NLI discussed in the preceding section, we tested a number of classifier and feature options, and most of our choices there are carried over to this work. In particular, we use the Liblinear SVM 1va (one versus all) classifier (Fan et al., 2008). Us-

[25] The work in this section is adapted from "Using other corpora in the 2013 NLI shared task" by Julian Brooke and Graeme Hirst, published in the *Proceedings of the 8th Workshop on Building Educational Applications Using NLP* (Brooke and Hirst, 2013c).

ing the TOEFL-11 corpus, we briefly tested the other options explored there (including SVM 1v1) as well as the logistic regression classifier included in Liblinear, and found that the SVM 1va classifier was still preferred (with our best feature set, see below), though the differences involved were marginal. Although small variations in the choice of C parameter within the SVM model did occasionally produce benefits (here and in our previous work), these were not consistent, whereas the default value of 1 showed consistently near optimal results. We used a binary feature representation, and then feature vectors were normalized to the unit circle. With respect to feature selection, our earlier work used a frequency cutoff of 5 for all features; we continue to use frequency cutoffs here; other common feature selection methods (e.g. use of information gain) were ineffective in our previous work, so we did not explore them in detail here.

With regards to the features themselves, our earlier work tested a fairly standard collection of distributional features, including function words, word n-grams (up to bigram), POS n-grams (up to trigram), character n-grams (up to trigram), dependencies, context-free productions, and 'mixed' POS/function n-grams (up to trigram), i.e. n-grams with all lexical words replaced with part of speech. Most of these had appeared in previous NLI work (Koppel et al., 2005; Wong and Dras, 2011; Wong et al., 2012), though until recently word n-grams had been avoided because of ICLE topic bias. Our best model used only two of these features, word n-grams and the mixed POS/function n-grams. This was our starting point for the present work. The Stanford parser (Klein and Manning, 2003) was used for POS tagging and parsing.

Obviously, the training set used varies throughout, and other differences in specific models built for each task will be mentioned as they become relevant. For evaluation here, we primarily use the test set for NLI shared task, though we employ some other evaluation corpora, as appropriate. During the preparation for the shared task, we made our decisions regarding models for two tasks with TOEFL-11 training according to the results in two training/test

Table 5.9: Feature testing for closed-training task, previously investigated features; best result is in bold.

Feature Set	Accuracy (%)
Word+mixed	76.8
Word+mixed+characters	72.0
Word+mixed+POS	76.6
Word+mixed+productions	77.9
Word+mixed+dependencies	**78.9**
Word+mixed+dep+prod	78.4

sets (800 per language for training, 100 per language for testing) sampled from the released training data. Since our research was focused on cross-corpus evaluation, we never created mechanisms for cross-validation in our system, and in fact it creates practical difficulties for the open-training task 2, so we do not include cross-validated results here.

5.5.3 Closed-training task

In the closed-training task, only the TOEFL-11 could be used as data. Our approach to this task primarily involved feature testing. Table 5.9 contains the results of testing our previously investigated features from Brooke and Hirst (2012d) (see preceding section) in the TOEFL-11, pivoted around the best set (word *n*-grams + mixed POS/Function *n*-grams) from that earlier work.

Some of the features we rejected in our previous work also underperform here, in particular character and POS *n*-grams. In fact, character *n*-grams had a much more negative effect on performance here than they had previously. Dependencies are clearly a useful feature in the TOEFL-11, this is fully consistent with our initial testing. CFG productions offer a small benefit on top of our base feature set, but are not useful when dependencies are also included, so we discarded them. Thus, our feature set going forward consists of word *n*-grams, mixed POS/function *n*-grams, and dependencies.

Next, we evaluate our feature frequency cutoff using this feature set (Table 5.10). We used

Table 5.10: Feature frequency cutoff testing for closed-training task; best result is in bold.

Cutoff	Accuracy (%)
At least 5 occurrences	78.9
At least 3 occurrences	79.5
At least 2 occurrences	79.7
All features	**80.2**

the rather high cutoff of 5 (for all features) in the previous work because of our much larger training set. We looked at higher values there, but for this task we focused on testing lower values.

Lowering our frequency cutoff is indeed beneficial, and we got our best result in the test set when we had no feature selection at all. This was not consistent with our preparatory testing, which showed some benefit to removing hapax legomena, though the difference was marginal. However, we did include a run with this option in our final submission, and so this last result represents our best performance on the closed-training task.

We tested several other feature options that were added to our system for this task. Inspired by Bykh and Meurers (2012), we first considered *n*-grams (up to trigrams) where at least one lexical word is abstracted to its POS, and at least one isn't (partial abstraction). Since dependencies were found to be a positive feature, we tried adding dependency chains, which combine two dependencies, i.e. three lexical words linked by two grammatical relations. We tested productions with wild cards, e.g. S $\rightarrow$ NP VP * matches any sentence production which starts with NP VP. Tree Substitution grammar fragments have been shown to be superior to CFG productions (Swanson and Charniak, 2012); we used raw Tree Substitution Grammar (TSG) fragments for the TOEFL-11[26] and tested a subset of those fragments which involved at least two levels of the grammar (i.e. those not already covered by *n*-grams or CFG productions).

Our final feature option requires slightly more explanation. Crossley and McNamara (2012)

[26]We thank Ben Swanson for letting us use his TSG fragments.

Table 5.11: Feature testing for closed-training task, new features; best result is in bold.

Feature Set	**Accuracy (%)**
Best	**80.2**
Best+partial abstraction	79.7
Best+dependency chains	78.6
Best+wild card productions	78.8
Best+TSG fragments	78.1
Best+MRC lexicon	54.2

report that metrics associated with word concreteness, imaginability, meaningfulness, and familiarity are useful for NLI; the metrics they use are derived from the MRC Psycholinguistic database (Coltheart, 1980), which assign values for each dimension to individual words. We used the scores in the MRC to get an average score for each dimension for each text, further normalized to the range 0–1; texts with no words in the dictionaries were assigned the average across the training set.

Table 5.11 indicates that all of these new features were, to varying degrees, a drag on our model. The strongly negative effect of the MRC lexicons is particularly surprising. We speculate that this might might be due partially to problems with combining a large number of binary features with a small number of continuous metrics directly in a single SVM. A meta-classifier might solve this problem, but we did not explore meta-classification for features here.

Finally, since that information was available to us, we tested creating sub-models segregated by topic and proficiency. The topic-segregated model consisted of 8 SVMs, one for each topic; accuracy of this model was quite low, only 67.3%. The proficiency-segregated model used two groups, high and low/medium (there were few low texts, so we did not think they would be sufficient by themselves for a viable model). Results were higher, 74.9%, but still well below the best unsegregated model.

Table 5.12: Number of tokens (in thousands) in external learner corpora, by L1.

L1	Corpus				
	Lang-8 (new)	**ICLE**	**FCE**	**ICCI**	**ICNALE**
Japanese	11694k	227k	33k	232k	199k
Chinese	7044k	552k	30k	243k	366k
Korean	5174k	0k	37k	0k	151k
French	536k	256k	61k	0k	0k
Spanish	861k	225k	83k	49k	0k
Italian	450k	251k	31k	0k	0k
German	331k	258k	29k	91k	0k
Turkish	51k	222k	22k	0k	0k
Arabic	218k	0k	0k	0k	0k
Hindi	11k	0k	0k	0k	0k
Telugu	2k	0k	0k	0k	0k

5.5.4 External corpora

We have already discussed most of the corpora used here earlier in Section 5.2; a summary of the number of tokens from each L1 background for each of the corpora is in Table 5.12. For our Lang-8 corpus we added more entries written since the first version was collected (58k on top of the existing 154k entries).

One obvious problem with using existing L2 corpora to classify L1s in the TOEFL-11 is the lack of Hindi and Telugu text, which we found were the two most easily confused L1s in the closed-training task. We explored a few methods to get data to fill this gap. First, we downloaded two collections of English language Indian news articles, one from a Hindi newspaper, the *Hindustan Times*, and one from a Telugu newspaper, the *Andhra Jyothy*. Second, we extracted a collection of English tweets from the WORLD twitter corpus (Han et al., 2012) that were geolocated in the Hindi and Telugu speaking areas; as with the Lang-8, these were combined to create texts of at least 250 tokens.[27] Our third Indian corpus consists of translations (by Google Translate) of Hindi and Telugu blogs from the ICWSM 2009 Spinn3r Dataset

[27] We extracted India regions 07 and 36 for Hindi, and 02 and 25 for Telegu; We can provide a list of tweet ids for reconstructing the corpus if desired. Our thanks to Bo Han and Paul Cook for helping us get these tweets.

Table 5.13: Number of tokens (in thousands) in Indian corpora, by expected L1.

L1	Indian Corpus		
	News	**Twitter**	**Blog**
Hindi	996k	146k	2089k
Telugu	998k	133k	76k

(Burton et al., 2009), which we used in other work on using L1 text for NLI, see Section 5.3. The number of tokens in each of these corpora are given in Table 5.13 .

5.5.5 Open-training task 2

In open-training task 2, other corpora could be used in addition to the TOEFL-11. Our approach to open-training task 2 is based on the assumption that in many ways it is a direct extension of the closed-training task. For example, we directly use the best feature set from that task, with no further testing. Based on the results in our initial testing, we used a feature frequency cutoff of 2 during our testing for open-training task 2; for consistency, we continue with that cutoff in this section.

We first attempted to integrate information from other corpora by using a meta-classifier, as was successfully used for features by Tetreault et al. (2012). Briefly, classifiers were trained on each major external corpus (including only the L1s in the TOEFL-11), and then tested on the TOEFL-11 training set; TOEFL-11 training was accomplished using 10-fold crossvalidation (by modifying the code for Liblinear crossvalidation to output margins). With the TOEFL-11 as the training set, the SVM margins from each 1va classifier (across all L1s and all corpora) were used as the feature input to the meta-classifier (also an SVM). In addition to Liblinear, we also output this meta-classification problem to WEKA format (Witten and Frank, 2005), and tested a number of other classifier options not available in Liblinear (e.g. Naïve Bayes, decision trees, random forests). In addition to (continuous) margins, we also tested using the classification directly. Ultimately, we came to the conclusion that any use of a meta-classifier

Table 5.14: Corpus testing for open-training task; best result is in bold.

Training Set	Accuracy (%)	
	no BA	with BA
TOEFL-11 only	79.7	79.2
+Lang-8	79.5	**80.5**
+ICLE	80.2	80.2
+FCE	79.6	79.3
+ICCI	77.3	76.7
+ICNALE	79.7	79.3
+Lang-8+ICLE	80.4	80.4
+all but ICCI	80.0	80.4

came with a cost (a minimum 2–3% drop in performance) that could not be fully overcome with the additional information from our external corpora. The result using SVM classifiers, margin features, and an SVM meta-classifier was 78.5%, well below the TOEFL-11–only baseline.

The other approach to using these external corpora is to add the data directly to the TOEFL-11 data and train a single classifier. For this, we use bias adaptation, which was introduced in the preceding section; briefly, it involves changing the bias (constant) factor of a model until the output of the model in some dataset is balanced across classes (or otherwise fits the expected distribution), addresses skewed results due to differences between training and testing corpora. In that earlier work, we used a separate development set, but here we rely on the test set itself; since the technique is unsupervised (in the 1va case), we do not need to know the classes. Table 5.14 shows model performance after adding various corpora to the training set (TOEFL-11 is always included), with and without bias adaptation (BA).

Many of the differences in Table 5.14 are modest, but there are are few points to be made. First, there is a small improvement using either the Lang-8 or the ICLE as additional data. The ICCI, on the other hand, has a clearly negative effect, perhaps because of the age or proficiency of the contributors to that corpus. Bias adaptation seems to help when the (messy and highly unbalanced) Lang-8 is involved (consistent with our previous work), but it does not seem useful

Table 5.15: Training set selection testing for open-training task 2; best result is in bold, best submitted run is in italics.

Training Set	Accuracy (%)	
	no BA	with BA
TOEFL-11 only	79.7	79.2
+Lang-8	79.5	80.5
+Lang-8 $r = 0.1$	81.4	81.6
+Lang-8 $r = 0.2$	80.6	81.5
+Lang-8 $r = 0.3$	81.0	80.6
+all but ICCI	80.0	80.4
+all but ICCI $r = 0.1$	81.5	**82.5**
+all but ICCI $r = 0.2$	81.0	*81.6*
+all but ICCI $r = 0.3$	80.9	81.3

applied to other corpora, at least not in this setting.

Our second domain adaptation technique involves training data selection, which has been used, for instance in cross-domain parsing (Plank and van Noord, 2011). The method used here is very simple: we count the number of times each word appears in a document in our test data, rank the texts in our training data according to the sum of counts (in the test data) each word that appears in a training texts, and throw away a certain number of low-ranked texts. For example, if a training text consists solely of the two words *I agree*[28] and *I* appears in 1053 texts in the test set, and *agree* appears in 325, then the value for that text is 1378. This method simultaneously penalizes short texts, those texts with low lexical diversity, and texts that do not use the same words as our test set. We use a fixed cutoff, r, which refers to the proportion of training data that is thrown away for each L1 (allowing this to work independent of L1 was not effective). We tested this on this method in tandem with bias adaptation on two corpus sets: The TOEFL-11 and the Lang-8, and all corpora except the ICCI. The results are in Table 5.15. The number in italics is the best run that we submitted.

Again, it is difficult to come to any firm conclusions when the differences are this small,

[28]This is not a made-up example; there is actually a text in the TOEFL-11 corpus like this.

but our best results involve all of the corpora (except the ICCI) and both adaptation techniques. Unfortunately, our initial testing suggested $r = 0.2$ was the better choice, so our official best result in this task (81.6%) is not the best result in this table. Performance clearly drops for $r > 0.2$. Nevertheless, nearly all the results in the table show clear improvement on our closed-training task model.

5.5.6 Open-training task 1

The central challenge of open-training task 1 was that the TOEFL-11 was completely off-limits, even for testing. Therefore, a discussion of how we prepared for this task is very distinct from a post hoc analysis of the best method once we allowed ourselves access to the TOEFL-11; we separate the two here. We did use the feature set (and frequency cutoff) from the closed-training (and open-training 2) task; it was close enough to the feature set from our earlier work (using the Lang-8, ICLE, and FCE) that it did not seem like cheating to preserve it.

Given our failure to create a meta-classifier in open-training task 2, we did not pursue that option here, focusing purely on adding corpora directly to a mixed training set. The central question was which corpora to add, and whether to use our domain-adaptation methods. Our experience with the ICCI in the open-training task 2 suggested that it might be worth leaving it (or perhaps other corpora) out, but could we come to that conclusion independently?

Our approach involved considering each external corpus as a test set, and seeing which other corpora were useful when included in the training set; corpora which were consistently useful would be included in the final set. Our original exploration involved looking at all of the corpora (as test sets), but it was haphazard; here, we present results just with the ICLE and the ICNALE, which are arguably the two closest corpora to the TOEFL-11 in terms of proficiency and genre. For this, we used a different selection of L1s, 12 for the ICLE, 7 for the ICNALE; all of these languages appeared in at least the Lang-8, and 2 of them (Chinese and Japanese)

Table 5.16: ICLE testing for open-training task 1; best result is in bold.

Training Set	Accuracy (%)	
	no BA	with BA
Lang-8	47.0	57.1
Lang-8+FCE	47.9	58.2
Lang-8+ICCI	46.4	54.8
Lang-8+ICNALE	46.9	57.5
Lang-8+ICNALE+FCE	47.7	**58.8**
Lang-8+ICNALE+FCE $r = 0.1$	46.6	58.2

Table 5.17: ICNALE testing for open-training task 1; best result is in bold.

Training Set	Accuracy	
	no BA	with BA
Lang-8	37.2	59.6
Lang-8+FCE	37.9	61.3
Lang-8+ICCI	35.7	61.4
Lang-8+ICLE	37.3	61.4
Lang-8+ICLE+FCE	37.6	61.7
Lang-8+ICLE+FCE $r = 0.1$	37.7	**61.9**

appeared in all corpora. Both sets were balanced by L1. Again, we report results with and without bias adaptation. The results for the ICLE are in Table 5.16.

The clearest result in Table 5.16 is the consistently positive effect of bias adaptation, at least 10 percentage points, which is line with our previous work. Adding both ICLE and ICNALE to the Lang-8 corpus gave a small boost in performance, but the effect of the ICCI was once again negative, as was the effect of our training set selection.

The ICNALE results in Table 5.17 support many of the conclusions that we reached in the ICLE (and other sets like the FCE and ICCI, which are not included here but gave similar results); the effect of bias adaptation is even more pronounced. Two differences: the slightly positive effect of training data selection and the positive effect of the ICCI, the latter of which we saw nowhere else. We speculate that this might be due to that fact that although the ICNALE is a college-level corpus, it is a corpus of Asian-language native speakers. Our theory is that

Table 5.18: Indian corpus testing for Open-training task 1; best result is in bold.

Training Set	Accuracy (%)	
	no BA	with BA
Indian news	50.0	54.0
Indian tweets	54.0	**56.0**
Indian blogs	51.5	**56.0**

Europeans are, on average, more proficient users of English (this is supported by, for instance, the testing from Granger et al. (2009)), and that therefore the European component of the low-proficiency ICCI actually interferes with using high proficiency as a way of distinguishing European L1s, a problem which would obviously not extend to an Asian-L1-only corpus. This is an interesting result, but we will not explore it further here. In any case, it would lead us to predict that including ICCI data would be a bad idea for TOEFL-11 testing.

Since we did not have any way to evaluate our Indian corpora (i.e. the news, twitter, and translated blogs) without using the TOEFL-11, we instead took advantage of the option to submit multiple runs, submitting runs which use each of the corpora, and combining the blogs and news.

We now offer some post-hoc analysis. With the TOEFL-11 data now visible to us, we first ask whether our specially collected Indian corpora can distinguish texts in the ICCI. The test set used in Table 5.18 contains only Hindi and Telugu texts. The results are quite modest (the guessing baseline is 50%), but suggest that all three corpora contain some information that distinguishes Hindi and Telugu, particularly if bias adaptation is used.

The results for a selection of models on the full set of TOEFL-11 languages is presented in Table 5.19. Since ours was the best-performing model in this task, we include results for both the TOEFL-11 training (including development set) and test set, to facilitate future comparison. Again, there is little doubt that bias adaptation is of huge benefit, though in fact our results in the Lang-8 alone, without bias adaptation, would have been enough to take first place in this

Table 5.19: 11-language testing on TOEFL-11 sets for open-training task 1; best result is in bold, best submitted run is in italics.

Training Set	Accuracy (%)			
	TOEFL-11 test		TOEFL-11 training	
	no BA	with BA	no BA	with BA
Lang-8	39.5	53.2	37.2	48.2
Lang-8+ICCI	36.9	51.0	34.9	46.3
Lang-8+FCE+ICLE+ICNALE	44.5	55.8	44.9	53.1
Lang-8+FCE+ICLE+ICNALE+Indian news	45.2	*56.5*	45.5	54.9
Lang-8+FCE+ICLE+ICNALE+Indian tweets	44.9	56.4	45.1	53.4
Lang-8+FCE+ICLE+ICNALE+Indian translated blog	45.4	50.1	45.7	49.9
Lang-8+FCE+ICLE+ICNALE+News+Tweets	45.2	57.5	45.5	55.2
Lang-8+FCE+ICLE+ICNALE+News+Tweets $r = 0.1$	44.9	**58.2**	45.0	**58.2**

task. Adding other corpora, including the Indian corpora but not the ICCI, did consistently improve performance, as suggested by our testing in other corpora. Although the translated blog data was useful in distinguishing Hindi from Telugu alone, it had an unpredictable effect in the main task, lowering bias-adapted performance. Training set selection does seem to have a small positive effect, though we did not see this consistently in our original testing.

5.5.7 Discussion

Our efforts in the 2013 NLI shared task focused on the potential benefits of external corpora. We have shown here that including training data from multiple corpora is effective at creating good cross-corpus NLI systems, particularly when domain adaptation, i.e. bias adaptation or training set selection, is also applied; we were the highest-performing group in open-training task 1 by a large margin. This approach can also be applied to improve performance even when training data from the same corpus is available, as in open-training task 2. However, in the closed-training task, despite testing a number of new features, we did not see much improvement on our simple model based on earlier work. Other teams clearly did find some ways to improve on this straightforward approach, and though there are some interesting insights

in various cases, the dominance of lexical features was a fairly clear pattern (Tetreault et al., 2013).

5.6 Investigating the effect of corpus variables on native language identification

5.6.1 Introduction

The work in the previous two sections has demonstrated the potential of a cross-corpus approach to native language identification. In this section,[29] we move from a focus on simply maximizing performance to a more exploratory look at how the specific qualities of NLI corpora influence cross-corpus performance. This is possible due in part to the recent plethora of new multi-L1 corpora, as mentioned at the end of Section 5.2. In that earlier section, we discussed some of the differences among the corpora; here, we introduce some simple qualitative metrics intended to reflect these differences. We and others (Tetreault et al., 2012) have already investigated to some degree the (positive, but ultimately diminishing) effects of corpus size when doing NLI; here, we control for this aspect and turn our attention to less easily quantified variables such as proficiency, genre, and topic diversity. Of the variables we are exploring, only proficiency has been addressed in previous NLI work, and only within the context of a single corpus, identifying whether high- or low-proficiency texts are easier to classify (Bestgen et al., 2012; Tetreault et al., 2012). Here, though, we are comparing the efficacy, as *training* data, of corpora which differ primarily with respect to one of these variables, showing that these variables all seem to have some effect on the resulting NLI classifier, albeit to varying degrees. The results presented here are important for NLI researchers working in more than one corpus, and could influence those collecting or expanding these corpora.

[29] This work was presented as "Investigating the influence of multi-L1 corpus learner variables on native language identification" by Julian Brooke and Graeme Hirst, at the 2013 Learner Corpus Research Conference.

5.6.2 Corpus metrics

In this section, we discuss some simple metrics which will be used to provide some concrete measurements of the somewhat intangible qualities such proficiency, genre, and topic diversity; in our results, we will present these numbers along with NLI accuracy. Note that in general these metrics are being calculated across the entire corpus; thus document boundaries are irrelevant.

Coleman-Liau Readability Index This is one of various popular readability metrics (Coleman and Liau, 1975); here, we regard it as a rough proxy for writing complexity. It is calculated using sentence length and word length, with a result which corresponds to American grade level.[30]

Type-token ratio A well-known measure of lexical diversity. Here, we intend to use it as a proxy for diversity of topic, though we note it might also be relevant to proficiency: highly proficient writers might have a larger vocabulary, whereas lower proficiency writers might make lots of spelling mistakes, both of which could boost TTR.[31]

Unigram entropy We consider the probability distribution over word types in the corpus, and calculate the entropy of that distribution. Higher values indicate that the probability mass is more evenly distributed among *n*-grams, which, like TTR, might indicate a diversity of topics.

Unigram KL-divergence across L1s As with unigram entropy, we calculate word-type probability distributions, but for each L1 in the corpus separately, and then average the KL-

[30] We choose this measure exactly because it uses word and sentence length, which are slightly more reliably calculated than other options. In general, these simple readability metrics, which also include Flesch (Kincaid et al., 1975), Dale-Chall (Dale and Chall, 1995), and FOG (Gunning, 1952) are highly correlated, with no major advantages to one over the other (van Oosten et al., 2010).

[31] Note that TTR cannot be reliably compared for texts of different length, but we are controlling for that here.

divergence (Kullback and Leibler, 1951) for all possible L1 pairings. Since KL-divergence is not well-defined for zero-probability situations, we first carry out add-one smoothing. This metric is intended to measure how (topically) different the texts in the individual L1s are.

Versine difference from testing corpus in LSA register space In Section 4.2, we showed that an LSA (i.e. bag-of-word, PCA) register space seemed to offer good differentiation of genres and interpretable dimensions at low k. Here, we use the BNC register space from that work. For each text in the training and testing corpora, we transfer the texts into the BNC register space, and then take the centroid of the individual text vectors to form a vector for the corpus. Then we take the cosine similarity between training and testing corpora, though for readability here (most of the corpora are very similar to each other in this space) we present the versine difference, which is simply one minus the cosine similarity. Our goal here is to get some measure of how close training and testing corpora are in terms of their genre/register.

5.6.3 Classification setup

For our classification experiments, we use the same SVM classifier as in the previous two sections. For these experiments, we test only two feature options: just mixed POS/word n-grams trigrams (delex), and both mixed POS/word n-grams and word n-grams (lex); in general, features are not the focus of this work, though we thought it would be interesting to see whether the exclusion of lexical features changed the story in some cases. Note that we are generally dealing with much smaller corpora here, so we do not expect lexical features to be as important in overall performance. We did not otherwise carry out any feature selection. In general, our counts across the different L1s for the training corpora are unequal, but recall that our classifier already accounts for that by using class weights. We also test both with or without the bias adaptation (BA) method; as with the shared task, we use the test set to change the bias.

We are comparing corpora by comparing their utility as training corpora in the NLI task. A very important step that we take here, which separates this work from all previous work in this area (that we are aware of) is to tightly control for overall token count as well as text length. Given a set of training corpora that we are comparing in the context of a testing corpus, we take the intersection of the L1s covered by all the corpora, and then, for each of those L1s, take the smallest token count for L1 among the training corpora, and use only that number of tokens for all the training corpora. Rather than using the original texts from the corpora, which vary widely in length,[32] we artificially combine and decompose texts of the same L1 to create new 'texts' that are almost exactly the average length of the texts in the testing corpus. The differences from the average are due to the fact that we respect sentence boundaries in all but the last text, which is potentially chopped in mid-sentence so the token counts are exactly the same across corpora. If the remainder from this text-creation procedure is less than half our target average, we just include it in the previous text, rather than creating a text that was potentially quite small. The texts in the testing corpus are preserved in the original form, but as we have before we equalize the counts across L1s. For both training and testing, texts are selected randomly to avoid potential ordering effects within the corpus.

Since the make-up of the training set determines the number of tokens used for all training sets and thereby influences both NLI accuracy as well as the metrics we are using to measure the corpora, comparisons of raw numbers will be possible only within a table, and not across tables. One serious drawback of the approach is that, for any given experiment (table) there is a necessary trade-off between the number of corpora included in the comparison and the number of languages included and tokens: if we try to compare all corpora, for example, we can do this only for Chinese and Japanese L1s, and only with the token counts of the smallest corpus (for these L1s), but, if we compare fewer corpora, we can have more L1s and more training tokens

[32] At one extreme, the average text length in the ICLE is over 500, whereas the average text length in the ICNALE is less than 100.

for each L1, which provides more reliability. Since each of these options has its advantages, we will mix them, as appropriate, though in general we always have at least three L1s.

5.6.4 Experiments

Most of our focus will be on comparisons of different corpora, but we begin by breaking down a single corpus, the TOEFL-11, into parts according to the essay quality and topic; the main goal is to get an initial sense of how our corpus metrics react to more controlled differences across 'corpora', and whether even these relatively small variations alone can have an effect on performance in a cross-corpus situation. For this, we use the ICLE for testing. The TOEFL-11 and ICLE overlap with 7 L1s: French, Spanish, Italian, Chinese, Japanese, German, and Turkish. These ICLE testing experiments included 249 texts per L1.

In the first experiment, we divide the corpus into 3 sub-corpora based on the provided proficiency annotation, i.e. high, medium, or low. We compare the results of training on each of these subcorpora as well as the original corpus, which includes an unequal mixture of all three proficiency levels (There are roughly three times as many high-proficiency texts as there are low-, and six times as many medium-proficiency texts as there are low-, though we took no steps to preserve these ratios in the subset used for training). For all training sets, the total number of tokens across all L1s was 147k.[33] The results are given in Table 5.20.

Looking first at NLI accuracy, we see once again the general pattern of improved performance when lexical features and bias adaptation are used. There is some volatility across conditions, but the general pattern seems to be better performance from medium proficiency text and/or the TOEFL-11 as a whole (which, as mentioned, is predominantly medium-proficiency texts); high- and low-proficiency texts tended to be worse. Note that Tetreault et al. (2012) showed that their general TOEFL-11 model did better on medium-proficiency TOEFL-11 texts

[33]Here and elsewhere we will omit the per-L1 breakdown, but we remind our readers that the number of training tokens is not balanced across L1s.

Table 5.20: 7-L1 native language identification in ICLE with corpus metric information for TOEFL-11 proficiency subsets. Delex.: Delexicalized *n*-grams, Lex: Lexicalized *n*-grams, BA: Bias adaptation, CLI: Coleman-Liau Readability Index, TTR: Type-Token Ratio, Ent.: Unigram entropy, L1-KL: Unigram KL-divergence across L1s, LSA diff.: Versine difference from testing corpora in LSA register space. Best results in column are in bold.

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11medium	35.4	**48.1**	39.8	**54.2**	8.9	0.057	8.8	0.51	0.002
TOEFL-11	**39.3**	47.4	**42.2**	53.0	9.2	0.059	8.9	0.44	0.001
TOEFL-11low	32.7	45.2	38.3	50.3	8.7	0.062	8.8	0.48	0.004
TOEFL-11high	36.4	42.9	39.4	47.9	9.6	0.058	9.0	0.44	0.001

than high- and low-proficiency texts, but this may have been due simply to the fact that the training set (the full TOEFL-11) involved more medium-proficiency texts. By contrast, the results here (controlled for total training set size) suggest that medium-proficiency texts might offer a better perspective on L1-influence for building a classifier. Of course, medium-proficiency is a relative notion here.

The right side of Table 5.20 shows the various metrics. For our interests here, the most important column is the first one: we do see that the proficiency of the text is directly reflected in our readability metric, with almost a full grade-level difference between low and high. It is worth noting that even the high-proficiency texts have lower CLI than the ICLE test set (9.9). Most of the other differences are more subtle, as we would expect, though it is worth noting that we see slightly higher TTR for low-proficiency texts, perhaps due to spelling errors, but the pattern is reversed for entropy. Lower-proficiency texts also showed more KL-divergence among L1s as well as a slightly higher LSA-derived difference from the test set, but in both cases these differences are small and, in the latter case, may be explained by the presence of essays which fail to meet the requirements of the genre. Note that none of the metrics directly predict the NLI accuracy in this case, but they do identify that the high- and low-proficiency

Table 5.21: 7-L1 native language identification performance in ICLE with corpus metric information for TOEFL-11 prompt subsets.

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11IP (avg)	38.4	**45.0**	41.5	**52.1**	9.2	0.054	8.5	0.41	0.003
TOEFL-11MP	**39.9**	44.5	**42.1**	50.3	9.3	0.061	8.9	0.43	0.001

texts relative to the others.

In the second experiment comparing subcorpora of the TOEFL-11, we break the corpus down by prompts. The TOEFL-11 has 8 prompts; rather than presenting the results for each of the 8, we present the average scores across all individual prompts (IP), and compare to the original (mixed-prompt) (MP) corpus. Here, the total token count for each training set is 136k. The results are in Table 5.21.

There are only small differences between mixed prompt and the averaged score for individual prompts with respect to performance in Table 5.21, though the less diverse training sets have a slight edge, which is the opposite of what we might have predicted. Importantly, our entropy and TTR metrics do clearly capture the difference between them. The individual prompts also have higher average LSA difference, though the difference is not as large as between high- and low-proficiency texts. We cannot come to any strong conclusions about the relationship between NLI and these variables based on this alone, but it does provide some initial intuitions, and also some useful information about the efficacy of our proxy metrics.

We continue testing in the ICLE corpus, though now we will widen our scope to include multiple training corpora. Since there is very little L1 overlap between the ICNALE and ICLE, we will not include the former in the discussion here. The ICLE and TOEFL are very similar with respect to the variables in question, while each of the FCE, ICCI, and Lang-8 are notably different in some respect. First, we compare the TOEFL-11 with the FCE, which is also a

Table 5.22: 7-language native language identification performance in ICLE with corpus metric information for FCE/TOEFL-11 comparison.

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	**47.5**	**56.2**	**55.0**	**66.5**	9.1	0.043	9.0	0.32	0.001
FCE	40.7	45.9	48.4	54.6	7.3	0.038	8.9	0.26	0.014

high-stakes examination but is of a different genre (mostly letters); we are using the same 7 L1s as in the previous section, and the total number of tokens for each training set is 302k. The results are in Table 5.22.

By comparison with the modest differences we saw when looking at sub-corpora within the TOEFL-11, here the contrasts are stark: a difference of over 10 percentage points when using bias adaptation, a result which is consistent for both lexicalized and delexicalized versions. Crucially, the results can be predicted by not one but two of our metrics: as compared to the TOEFL-11, the FCE has a notably higher LSA register difference, and it has a clearly lower CLI (i.e. higher readability) as compared to both the TOEFL and the ICLE (which are both over 9). The former is exactly what we would predict under these circumstances, while the latter is a little more difficult to interpret: The FCE is an upper intermediate test, and we might expect the proficiency of students to be roughly comparable to those taking the TOEFL. The other possibility is that the difference in genre is having a direct influence on the CLI metric, perhaps because letters tend to involve shorter sentences than essays.

Next, we compare the TOEFL-11 with the Lang-8 corpus. Like the FCE, the Lang-8 is also of a different genre, and here we do expect lower proficiency levels: many users of the site are clearly beginners, though there are advanced learners as well. For this experiment, we use 1.9 million tokens for training; the results are in Table 5.23.

The results are similar, though more pronounced than with the FCE. Like the FCE, the

Table 5.23: 7-L1 native language identification performance in ICLE with corpus metric information for Lang-8/TOEFL-11 comparison.

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	**58.9**	**65.3**	**70.2**	**75.4**	9.1	0.020	9.0	0.31	0.001
Lang-8	47.8	53.8	54.1	62.3	7.1	0.034	9.9	0.51	0.010

Table 5.24: 4-L1 native language identification performance in ICLE with corpus metric information for ICCI/TOEFL-11 comparison.

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	**70.4**	**76.8**	**73.7**	**82.0**	9.2	0.032	9.0	0.32	0.001
ICCI	52.7	58.4	52.3	59.6	5.0	0.030	8.4	1.33	0.023

Lang-8 corpus has markedly lower CLI (more readable) and a large LSA difference, though less than the FCE. Unlike the FCE, it also has notably higher TTR, entropy, and cross-L1 KL-divergence, suggesting in general a much more diverse, less-controlled dataset. In this case, though, this does not result in good performance relative to a corpus that is strongly similar to the testing corpus.

Finally we look at the ICCI corpus, which is also an essay corpus but was collected from grade school rather than college-level students; many of the essays are also descriptive, rather than argumentative as with the ICLE and TOEFL-11. From this point on, we use only 4 L1s: Chinese, Japanese, Spanish, and German, though the number of texts per L1 in the testing set remains the same. The number of tokens used here is 616k. The results are in Table 5.24.

The performance gap between the ICCI and TOEFL corpus is the largest seen so far. Several of the metrics indicate potential problems: first, according to the CLI, the corpus is another two grade levels simpler than the FCE and Lang-8. The LSA difference is once again perfectly

Table 5.25: 4-L1 native language identification performance in ICLE with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	**66.3**	**76.7**	**67.7**	**80.4**	9.0	0.059	8.9	0.33	0.001
FCE	56.6	61.7	61.9	69.2	7.4	0.054	8.8	0.27	0.012
Lang-8	41.1	53.2	40.5	58.6	7.6	0.086	9.6	0.50	0.006
ICCI	49.0	57.5	48.1	56.9	5.3	0.062	8.6	1.13	0.024

aligned with the variation in CLI; thus far, the two are almost indistinguishable. We also see a much lower entropy in the ICCI, which is not surprising given its low complexity. The most striking result, though, is the much higher KL-divergence across L1s: this indicates that there are major differences in the words used across the L1s, which might also be the cause of the relatively poor performance, though we note that the ICCI is still doing well above chance (25%).

Table 5.25 combines all 4 training corpora in a single comparison, with the 4 L1s. There are only 142k tokens for each training set. Again, the TOEFL-11 is the clearly superior training set for the ICLE, followed by the FCE. In most of the conditions, the Lang-8 is actually the worst of the 4, though it marginally beats out the ICCI with lexical features and bias adaptation. Based on the LSA register difference, though, we would expect the Lang-8 to do much better; one potential explanation, one which is consistent with the TOEFL-11 topic sub-corpora results, is that the higher diversity of the Lang-8 is not a positive when the training set is so small. Based on the LSA register difference, there is no evidence here that the ICCI is particularly close to the ICLE, despite both being 'essay' corpora, and the differences in learner proficiency and/or the distribution of topics across L1s may all be playing a role in the inferior preformance relative to the other corpora here.

The next experiment involves the same 5 corpora (and 4 L1), but in this experiment the

Table 5.26: 4-L1 native language identification performance in TOEFL-11 with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
FCE	54.6	**58.9**	**59.6**	**63.1**	7.4	0.054	8.8	0.27	0.008
ICLE	**54.7**	56.1	56.8	59.8	9.5	0.076	9.4	0.82	0.001
Lang-8	39.0	53.5	38.4	57.6	7.6	0.086	9.6	0.50	0.003
ICCI	41.8	48.8	42.6	50.3	5.3	0.062	8.6	1.13	0.016

TOEFL-11 is the testing corpus. Given the strong preference for TOEFL-11 training in the ICLE, we might expect the ICLE to similarly dominate the competition in the TOEFL-11. The results, which use 142k tokens for training and 1100 texts per L1 for testing, are in Table 5.26.

Though the margin is not large, the FCE is a better training set than the ICLE for most of the conditions. The likely answer to why the ICLE is a less-than-ideal training set, despite being so close to the TOEFL with respect to genre and proficiency, is visible in its cross-L1 KL-divergence: though not as high as the ICCI, it is markedly higher than the other corpora, including the test corpus (which has a KL-divergence of just 0.33, closest to the FCE). Again, the Lang-8 seems to be underperforming relative to its LSA difference from the test set. Another striking aspect of the table is the effect of bias adaptation. Note that all the corpora benefit from bias adaptation, but the degree to which they benefit varies widely, a phenomenon that did not occur to this degree when testing in the ICLE. Both these results underline that there are major differences between ICLE and TOEFL-11, which are otherwise 'close' corpora.

Our next test set is the FCE, though our first experiment uses the same set of L1s and corpora as above. Since the FCE is the smallest corpus in the study (at least on number of texts per L1 basis), using it as a test corpus involves a major drop in the size of the testing text, to 66 per L1, but a major increase in the number of tokens used, to 610k. The results for FCE testing for these 4 L1s are in Table 5.27.

Table 5.27: 4-L1 native language identification performance in FCE with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
Lang-8	**56.4**	**59.5**	**60.6**	**69.3**	7.3	0.047	9.7	0.52	0.001
TOEFL-11	54.9	59.1	56.4	65.5	9.2	0.032	9.0	0.32	0.008
ICLE	39.4	47.7	43.2	57.2	9.9	0.035	9.4	0.93	0.010
ICCI	30.7	45.8	27.7	46.6	5.0	0.030	8.4	1.33	0.002

In this experiment, the Lang-8 comes out on top by a substantial margin over the TOEFL-11, which is in turn much better than the ICLE. The ICCI is again bringing up the rear. Again, there are a number of possible explanations for the good Lang-8 performance, most obviously the relatively close correspondence in proficiency and genre, as measured by the CLI and LSA difference. These are so clearly in tandem that we must question whether we are really measuring two separate quantities, though it is worth noting that the large difference in CLI between the ICCI and FCE does not result in a large LSA register difference, as it does for the more proficient ICLE and TOEFL-11. Another possible explanation is, of course, the higher entropy and TTR, possibly in combination with the much larger training size in this experiment. The KL-divergence among L1s is also a reasonably good predictor of performance. Bias adaptation is again important across all training corpora, and lexical features offer a major gain, but only for corpora with lower cross-L1 KL-divergence.

If the Lang-8 does better here (relative to its depressed performance in earlier experiments) in part because increased amounts of training data are important for building good models from diverse corpora, we can confirm this by simply limiting the amount of training data. The results of this experiment, identical to Table 5.27 except for the fact that we are using only one-fourth of the training data, is in Table 5.28. As predicted, the TOEFL-11 now has the advantage, and so we can conclude that overall corpus diversity does interact with the quantity of data used:

Table 5.28: 4-L1 native language identification performance in FCE with corpus metric information for various training corpora with less data than in Table 5.27

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	37.9	**52.3**	**43.9**	**56.4**	9.2	0.056	8.9	0.40	0.008
Lang-8	**38.6**	49.6	39.0	53.4	7.4	0.084	9.5	0.59	0.001
ICLE	28.0	40.9	30.3	46.2	9.8	0.066	9.3	0.87	0.010
ICCI	32.2	41.3	29.2	42.4	5.0	0.049	8.3	1.17	0.002

diverse corpora need more data. We also note that in this experiment, the difference between lexical and non-lexical features is much less pronounced than in the previous experiment. This indicates that lexical features benefit greatly from increased data.

The next experiment also involves testing in the FCE, but compares the Lang-8, TOEFL-11, and the as-yet-unused ICNALE in a different set of L1s: Japanese, Chinese, and Korean. The number of tokens used is 713k. See the results in Table 5.29. The ICNALE has a similar CLI as the TOEFL-11 (and the ICLE), and the performance in the FCE is similar as well. It has much lower TTR and entropy (the lowest seen so far), but this does not seem to affect performance in the (similarly restricted) FCE. Both the KL-divergence among L1s and the LSA difference also correctly predict the ordering of training corpora when lexical feature and bias adaptation are used, though without bias adaptation the ordering is exactly opposite, and for the first time bias adaptation actually worsens performance (in the TOEFL-11), which we have previously seen only when applying bias adaptation when both testing and training sets are from the same corpus.

Table 5.30 shows the results when the ICNALE is used as a test corpus, with the same languages and corpora as the previous experiment. The test set is 600 texts per L1, with a total of 100k training tokens. The results are straightforward: the LSA distance between the TOEFL-11 and the ICNALE, both essay corpora, approaches zero (0.0003), and the TOEFL-

Table 5.29: 3-L1 (Asian) native language identification performance in FCE with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
Lang-8	**51.0**	**55.6**	49.0	**60.6**	7.1	0.041	9.6	0.42	0.001
ICNALE	45.5	50.0	47.5	54.0	9.2	0.017	8.5	0.29	0.004
TOEFL-11	46.5	50.5	**54.0**	52.5	9.3	0.030	9.0	0.18	0.007

Table 5.30: 3-L1 (Asian) native language identification performance in ICNALE with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	**61.6**	**63.4**	**69.2**	**70.3**	9.3	0.068	8.8	0.27	0.000
Lang-8	49.6	51.2	52.7	55.0	7.3	0.087	9.3	0.51	0.003
FCE	48.9	49.5	51.9	52.3	7.7	0.058	8.7	0.20	0.003

11 is markedly preferred over the Lang-8 and FCE, which are roughly similar in performance, exactly what we would expect based on the LSA distance. The effect of bias adaptation is quite small relative to what we've seen before.

Table 5.31 contains testing in Lang-8 for the 3 Asian L1s. The test set is 288 texts per L1, with only 100k training tokens. Here the three corpora are fairly similar, though the ICNALE has an edge over the other two. This is somewhat surprising, given what we have seen so far, since the FCE appears closer via LSA distance (as well as CLI), and we would not expect either lower entropy or high KL-divergence among L1s to result in higher performance. However, it is worth noting that the KL-divergence of the Lang-8 is actually quite a bit higher than any of the corpora being compared here (0.51). If we consider the possibility that relative (that is, relative to the test corpus) rather than absolute cross-L1 divergence is playing the key role here, then the high performance of the ICNALE relative to the FCE makes sense. If we return to the

Table 5.31: 3-L1 (Asian) native language identification performance in Lang-8 with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
ICNALE	**42.1**	**43.9**	**48.3**	**48.4**	9.2	0.049	8.5	0.31	0.002
TOEFL	39.7	41.2	46.3	47.6	9.3	0.068	8.8	0.27	0.004
FCE	37.8	38.0	43.2	43.4	7.7	0.058	8.7	0.20	0.001

Table 5.32: 4-L1 native language identification performance in Lang-8 with corpus metric information

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFL-11	**42.0**	**45.5**	**47.0**	**51.0**	9.2	0.032	9.0	0.32	0.003
ICCI	37.6	37.8	38.9	42.9	5.0	0.030	8.4	1.33	0.002
ICLE	35.4	39.8	37.9	42.6	9.9	0.035	9.4	0.93	0.004

4-L1 task (Japanese, Chinese, German, Spanish) and compare the TOEFL-11 with the ICCI and ICLE, we see that the ICCI and ICLE are markedly poorer training sets for the Lang-8, either due to the much higher KL-divergence among L1s or, alternatively, the large relative differences between these corpora and the training set (which has a KL-divergence of just 0.32). For both Lang-8 experiments, the effect of bias adaptation is low. These results are in Table 5.32. The number of tokens used is 610k.

Our final two experiments use the ICCI as testing corpus; there are 612 testing texts per language, and, for the first experiment, 176k tokens for training. The results in Table 5.33 show that the ICLE is the best training corpus for the ICCI. Although we might predict that this is due to both corpora being of same genre, everything we have seen thus far suggests that the ICCI bears relatively little resemblance to other essay corpora, and the difference in CLI between these two corpora is extreme. The explanatory commonality is that both corpora have

Table 5.33: 4-L1 native language identification performance in ICCI with corpus metric information for various training corpora

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
ICLE	40.3	**52.2**	41.8	**56.4**	9.5	0.070	9.4	0.83	0.011
TOEFL-11	43.6	44.0	**48.0**	49.5	8.9	0.054	8.9	0.34	0.009
FCE	**44.8**	45.3	47.8	47.3	7.2	0.049	8.9	0.28	0.001
Lang-8	37.5	36.4	39.7	38.6	7.5	0.079	9.6	0.5	0.003

Table 5.34: 4-L1 native language identification performance in ICCI with corpus metric information for TOEFL-11 proficiency subsets

Corpus	NLI Accuracy (%)				Corpus Metrics				
	Delex.		Lex.		CLI	TTR	Ent.	L1-KL	LSA Diff.
	no BA	w/BA	no BA	w/BA					
TOEFLlow	**43.1**	**49.2**	**43.4**	**49.8**	8.6	0.074	8.7	0.53	0.007
TOEFL	33.5	37.1	33.1	38.5	9.1	0.073	8.8	0.53	0.009
TOEFLmedium	33.0	34.5	33.0	33.7	8.9	0.071	8.8	0.64	0.009
TOEFLhigh	31.9	28.0	32.5	29.0	9.5	0.072	8.9	0.53	0.012

a very high cross-L1 KL-divergence, which we had originally assumed to be a categorically negative property for a L1 training corpus to have; this result as well as those above suggest strongly otherwise. Interestingly, the high performance of the ICLE relies on bias adaptation, whereas in other corpora bias adaptation has little effect.

Finally, the ICCI can offer a new perspective on our first comparison: i.e. the proficiency levels in the TOEFL-11 corpus. Table 5.34 has those results. With a lower proficiency testing corpus, lower proficiency training texts are obviously much preferred. Here, the classifier built from the high-proficiency TOEFL-11 texts performs only slightly better than chance.

5.6.5 Discussion

In this section, we have presented the most thorough investigation of multi-L1 learner corpora to date, and in doing so we have connected the task of native language identification to some of the more general interests of this thesis. As a stylistic task, NLI is particularly susceptible to the effects of other kinds of stylistic variation, and the differences we have seen in terms of the usefulness of corpora for training NLI classifiers underlines this effect. Other than the ICCI, every single corpus was found to be the best training corpus in at least one experiment (when corpus and text length are controlled for). One obvious conclusion from what we've seen here is that the relationships between NLI and corpus variables are fairly complex, and we should not expect there to be one best corpus that will result in the best classifier for all situations. This is a key point, because although most if not all NLI research has focused on high-proficiency essay corpora with a small set of topics, many of the potential applications of NLI would not involve texts like these. One potential use of this research, with respect to improving performance in NLI, is that the corpus metrics here could be used to filter a large database of texts for more targeted training even when only general properties of the input texts are known. For researchers interested more in understanding the linguistic phenomena, this work should similarly highlight the limits of working in a restricted set of texts: if a generalizable NLI classifier cannot be built from a corpus like the ICLE, can a generalizable theory of language transfer be derived from just the ICLE?

Looking at the individual variables, there are some clear effects related to genre and proficiency, for instance a preference for the TOEFL-11 when testing in other high-proficiency essay corpora (i.e. the ICLE and the ICNALE) over corpora that differ somewhat in genre and proficiency (the FCE, Lang-8, and ICCI). Interestingly, the FCE and Lang-8 seem to be relatively close corpora according to our genre and proficiency metrics, which is not immediately clear based on our intuitions about these corpora: why should (often narrative) journal entries

be more like (interactive) letters than like argumentative essays? More generally, our results cannot say conclusively whether genre or proficiency is more important, though the universally poor performance of the ICCI and the preference for the FCE as the training corpus in the TOEFL-11 do suggest that proficiency level might be a more important factor than genre; these results might be equally attributable to topic diversity effects, however. Our results with proficiency sub-corpora of the TOEFL-11 also indicate a sizeable effect due to proficiency that is seemingly independent of genre differences, though even there we cannot be fully certain since conforming to the conventions of the genre is an important mark of successful writing. In general, these two variables are very difficult to tease apart, as reflected in the fact that our proficiency-focused complexity measure and the genre-focused LSA difference were more or less in lock-step, offering equally correct predictions in many cases. Given that we've already seen a certain amount of redundancy between traditional text statistics and lexicon-based measures, i.e. our lexical readability work (Section 3.5) and our comparison between LSA and multidimensional analysis (Section 4.2), and, more generally, the inherent challenges of distinguishing different stylistic dimensions that can be roughly understood in terms of the oral/situational vs. written/cultural dichotomy (Section 3.6), this is of no great surprise. Additional corpora, better annotation of existing corpora, or more targeted metrics might be needed to say anything definitive about these variables independently.

Our original hypothesis was that overall corpus diversity in the training set would have a positive effect in our cross-corpus evaluation. However, we see clearly that, in the context of a very restricted training set size, diversity appears to be a liability, explaining the underperformance of the Lang-8 in many of these experiments relative to what we would predict based on the (otherwise fairly reliable) LSA difference metric and our promising results using the Lang-8 in earlier sections of this thesis. We also note that sheer diversity is unlikely to overcome the influence of other variables; our direct comparison of the Lang-8 with the TOEFL-11

(in the similar ICLE) uses fairly large amount of data (1.9 million tokens), but the TOEFL-11 wins out by a significant margin. However, we would need larger corpora to make this point conclusively.

With regards to topic diversity across L1s, we assumed that such variation almost always result in a less generalizable classifier, particularly with (though, given our results in Section 5.4.5, not limited to) lexical features. Our results here do not fully support this, however. It is generally true that the high L1 diversity ICCI and ICLE do poorly (or at least, more poorly than would be otherwise expected) in most experiments, but in two cases (testing in the Lang-8 and the ICCI), it is the *lower* cross-L1 diversity corpora that seem to be at a disadvantage. A hypothesis that fits our results somewhat better is that it is the relative rather than absolute cross-L1 topic difference (that is, relative to the test set) that is the main determining factor for NLI performance. This would explain, for instance, the anomalous (compared to our results as well as those of Tetreault et al. (2012)) cross-corpus results reported by Bykh and Meurers (2012), who trained in the ICLE but for testing used a 'corpus' that was actually an amalgamation of several independently-collected essay corpora (one for each L1); such a 'corpus' would almost certainly have high L1 diversity, and therefore, under this hypothesis, the ICLE might very well be a good training corpus for such a test set. However, we have a good reason to be somewhat skeptical of this conclusion as well, since if relative L1 topic diversity were the sole determinant, we would expect the TOEFL-11 and ICLE to be symmetric, which they are clearly not. More generally, the fact that the ICLE and ICCI are so much more extreme than the only fully spontaneous corpus included (the Lang-8; all the other corpora use writing based on prompts) suggests there is something particularly artificial about these corpora, and we are still better off avoiding them if possible, or at least taking steps to mitigate the effect of this unwanted diversity.

Though lexical features generally provide a boost to performance, particularly when more

data is available, the more important point coming from our inclusion of the lexical/delexicalized distinction among features is that most of the results are independent of these distinctions: including lexical features does not generally change the nature of the underlying variation. There are, however, fairly extreme differences in the effect of bias adaptation across our experiments: one clear tendency is that bias adaptation is more important when the average text lengths are relatively high (recall that our training sets are built to have the same lengths as their test corpus), e.g. the ICLE. The testing corpora for which bias adaptation has generally little effect (the Lang-8, the ICCI, and the ICNALE) are (on average) the three corpora with the shortest text lengths. The fact that bias adaptation is more important with longer texts does not alone explain everything we see here. As with the other variables above, we have identified at least one key pattern in the data, but there is more going on than our experiments here can elucidate.

Chapter 6

Conclusion

Style, defined broadly as I have done here, is a pervasive quality of language. And yet it is one that is routinely ignored or minimized in the field of computational linguistics. This is possible in part because the field as a whole has a tendency towards a focus on one particular corpus, e.g. the Wall Street Journal, as the standard for a certain task, with research mostly limited to optimizing performance in that one corpus. Though often successful for controlling the effects of style, this myopic view to solving CL tasks is troubling for many reasons; see our discussion of native language identification and the ICLE corpus; or the problems with the PAN intrinsic plagiarism shared task. There may, of course be instances when a task is focused on highly restricted genre (e.g. the medical domain) and a single, stylistically monotone corpus can be representative, but for many other tasks this is a rather naive assumption, with the result being that a researcher may spend years chasing particular eccentricities of a corpus rather the broader patterns that persist across text types.

Perhaps even more troubling than the idea that one single corpus can ever represent the variety of language is the idea that corpora are infinitely variable, that there are no general patterns to be captured and no way that information from one corpus can help us with another. The work presented here is, I hope, a strong rebuttal to this conception of text variation. Throughout this thesis, I have shown how patterns *do* transfer: how various corpora show the same basic

stylistic patterns, how written texts of a 'single' genre (blogs) can be leveraged for identifying variation in texts as diverse as informal interviews and literature, how a diverse collection of sources for non-native style can be applied to the task of native language identification. If we conceive of style as a set of text types, then it is almost certain that this is a very large set indeed. But language is, perversely, simpler than that: there are broad patterns, and we have touched on many of them here. And, if we go even deeper than some of the words we have used here to describe stylistic variation (formality, readability, age, proficiency), we can see the influence of education and culture as creating a fundamental polarity in language. The fact of human learning is so universal that we should not be concerned that these stylistic patterns are mere artifacts of any particular corpus; rather, we should be concerned that in CL we are solving the same problems again and again, each time under a slightly different guise.

In Chapter 2, we looked at several perspectives in linguistics that deal explicitly with stylistic phenomena. Although distinct, there are interesting commonalities across these approaches: for instance, framing in terms of polarity and spectrum is common, and the kinds of variation addressed (e.g. argumentativeness, subjectivity) are clearly related if not fully identical. Though lexical features are central to prescriptivism, this is much less true in descriptive linguistics; in my opinion, this is *not* because the lexicon is not relevant to the kinds of variation discussed here, but rather that linguistics has a long-standing disinterest in 'messy' lexical matters, one that extends to its most prominent theorists (e.g. Chomsky). It is rather unfortunate that computational linguistics has inherited this bias against the lexicon in the context of style; with our techniques to identify relevant patterns in huge text corpora, I believe we are in a much better position to address the stylistic lexicon than linguists or even lexicographers.

The most important result in Chapter 3 is the demonstrated effectiveness of topic-based approaches, i.e. LSA and LDA, as applied to the derivation of lexical information relevant to the stylistic dimensions suggested in Chapter 2. This approach is a direct refutation of the idea

that topic is lexical and style is not, and offers resources for the work in the following chapter. Our initial work in induction focused on a single stylistic dimension, formality, but we later expanded this to include several more aspects of style when it became increasingly clear that looking at stylistic dimensions independently is problematic; this is at least in part attributable to the notion of a main cline of register (style) (Leckie-Tarry, 1995), as discussed in Chapter 2. Another important result that carries over into the work in other chapters is the idea that important stylistic variation is contained within the first few dimensions of our blog-derived, LSA-reduced lexical vectors. This leads us directly to the conclusion that style is a primary cause of lexical variation in language.

One idea that appears first in Chapter 3 is that term frequency is less relevant to style: we build better lexicons when we use binary feature representations. Not only does this distinguish style from topic (where the number of times a word appears in a text is key to its topicality), it is also a significant departure from traditional approaches to style and register, which are also often based on frequency (of function words, classes of words, etc.). Another contrast with topic is that, with some exceptions like sub-genres in literature, style is mostly determined by contextual factors that operate at the level of the text (or higher), and therefore there are also fewer locality effects, and, by extension, more context that can be used reliably for co-occurrence. If we consider pushing further with respect to differences between topic and style, we could integrate the Leckie-Tarry conception of style directly into our Bayesian model, though, given our negative results with the correlated topic model, we are somewhat skeptical that this would be effective. More promising, but also more difficult, would be to build a probabilistic model which views style as a choice among synonyms after the semantic content has been chosen, rather than a choice from the entire vocabulary, similar to the restrictions placed on, for instance, variables in the field of sociolinguistics.

We showed that going beyond corpus co-occurrence when building our formality lexicon

was a promising approach, as was a holistic, multi-style approach that used (semi-)supervised refinement of raw scores; we have not, however, combined the two ideas here. It would be particularly interesting to see if we could use features specific to each style while at the same time preserving the interaction between styles. The inclusion of the notion of sense could help improve the lexicon, and it would be interesting to try to derive senses and styles simultaneously. Similarly, multi-word extensions of our lexicons are a natural and reasonably straightforward next step, and could be quite useful for our various applications; when we look at the annotations in *To The Lighthouse*, for instance, we see that many of the style-indicating elements are really expressions, not individual words. It might be useful to approach this from a fully syntactic perspective, looking at lexicogrammatical patterns rather than simply *n*-grams. Also, the 6-style lexicon covers a fairly broad range of stylistic phenomena, but there are other possible styles that might fit into our framework here; one example is specificity. We could also consider breaking down our existing styles into their constituent (sub-)styles, for instance profanity is a kind of colloquial language, dynamism (i.e. motion-related) is kind of concrete language, etc.

Much of the work in Chapter 3 is quite recent, but we note that our approach to formality has influenced a small group of researchers interested in this specific stylistic dimension (Lahiri et al., 2011; Mosquera and Moreda, 2012; Sheika and Inkpen, 2012). The work on extracting a lexicon of potential sociolinguistic variables contributed to a recent discussion in linguistics related to data mining as an approach to identifying patterns for linguistic study, and as such was featured on the popular linguistics-themed weblog Language Log.[1]

As a collection of stylistic tasks, Chapter 4 is necessarily scattered and incomplete; as seen in our survey (which in itself is not by any means exhaustive) there are many more relevant stylistic tasks than I can reasonably address in the context of this thesis, and of course the more that are addressed, the more diverse this section becomes. That said, the contributions included

[1]Mark Liberman, Corpus-Wide Association Studies, March 11, 2012; http://languagelog.ldc.upenn.edu/nll/?p=3833

here, as a whole, make the key connection between the lexical stylistic information derived from corpora in Chapter 3, and tasks included under the three main stylistic facets introduced in Chapter 2. For example, the genre differentiation task, the only one in this chapter where no style-specific lexicon is applied, makes a clear point about the relative merit of lexical versus non-lexical features in the context of genre: a lexical approach is more straightforward, easily extended to more languages, quantitatively superior, and qualitatively comparable to the classic approach of Biber. In distinguishing sociolinguistic factors in a spoken corpus using our formality lexicon, we show that our aesthetic conceptions of style are relevant to social variation, in ways that we would expect but are nonetheless non-trivial, especially given the distinction between corpora. Clipping prediction is admittedly a rather superficial task, but it specifically narrows in on style as independent from topic, and the good performance of our formality lexicon relative to a more sophisticated (though also lexical) supervised approach is an important result, as is preference of the supervised model for the style-specific parameters we saw during lexical induction.

Literature is in general an excellent testing ground for stylistic resources, since a great deal of stylistic variation is usually expected, both across texts and often within texts, and in literature we see a coming together of all our stylistic interests: the aesthetic goals of the writer, the social background of his or her characters, and, in the case of novels, the sub-genres which reflect the functional requirements of the genre. Our extrinsic lexical resources (some from Chapter 3, some from other sources) form a key part of our unsupervised approach to distinguishing and then clustering voices in *The Waste Land*; not only do lexical features compare favorably with non-lexical features in this context, we also establish the basic potential of lexical information derived independently from large corpora as compared to the distribution of surface lexical features. Our small study applying the 6-style model to the novel *To The Lighthouse* demonstrates another connection between aesthetic lexicons and genre tasks, but more

importantly it highlights an advantage of human-interpretable lexicons beyond the utility in any particular task: human interpretability facilitates qualitative evaluation, analysis, and productive interaction with researchers in related disciplines. Finally, the problematic evaluation at PAN, using novels in the intrinsic plagiarism task, is an clear example of why the connections between the different facets of style need to be better recognized by the computational stylistics community.

With respect to tasks addressed in Chapter 4, it would be interesting to make connections between performance on these tasks and the intrinsic quality of lexicons derived in Chapter 3; as it stands now, we cannot be certain that our efforts to improve the lexicons are having any effect in downstream tasks. In the genre differentiation work (which does not use a lexicon), making an explicit comparison between dimensions we qualitatively identified based on genre distributions and the contents of the corresponding lexicon from Chapter 3 would strengthen that our conclusions, as well as give us a sense of how important it is to go beyond the dimensions offered by LSA. The work with sociolinguistic factors and novel sub-genres are here just simple experiments to showcase our formality and 6-style lexicons, respectively, but they could be expanded into full tasks, where we test a wide range of features. In the latter work, we are also interested in using our student annotation in a way which accepts the possibility of multiple valid interpretations, and we would like to test our conclusions relevant to free indirect discourse in other modernist novels which contain this sub-genre.

In addition to the tasks mentioned in our survey of relevant work, it is likely that the information we have derived could be useful for word sense disambiguation (see, for example, Turney et al. (2011)). Multiple senses are particularly common at the colloquial end of the stylistic spectrum, and this is an area which has not, to our knowledge, gotten much attention in the field. Also, a very recent popular task that it would be interesting to address given its obvious stylistic nature (and, like native language identification, the confounding effect of topic)

is the detection of fake online reviews (Ott et al., 2011).

The work on native language identification in Chapter 5 is clearly distinct from all the other work here, since NLI is a very different kind of stylistic task; for instance, we must relax our admitted preference for general human interpretability, since only multilingual speakers really have access to the kind of lexical information needed to make judgements in these cases. We also use very different methods; with the exception of Section 5.3—which is very much in the spirit of our lexical induction from Chapter 2, relying as it does on essentially inexhaustible L1 web data—we focus mainly on supervised approaches which build 'lexicons' only in the sense that the feature weights of the model reflect lexical preference. However, our diverse corpora and novel approach to evaluation allows us to be much more confident that we *are* capturing 'true' language transfer than if we relied on cross-validation in a single corpus; several other recent papers in NLI have adopted our cross-corpus evaluation as a way to demonstrate that their results are getting at real L1 transfer (Bykh and Meurers, 2012; Tetreault et al., 2012; Swanson and Charniak, 2013). There are more topical aspects to NLI, i.e. place names, but we definitively show that the benefit to lexical features goes far beyond that, a result that is consistent with the results of the first NLI shared task (Tetreault et al., 2013). We also demonstrate that not only is lexical variation relevant to this stylistic task, but that traditionally stylistic features (e.g. POS) are not immune from effects from topic: in short, this artificial delimitation of style and topic based on feature type is simply untenable.

Although the NLI work is mostly parallel to the other stylistic work here, there are at least three points of intersection. The first is that other stylistic variables (e.g. 'proficiency') have a major effect on NLI, which is made explicit here. Second, an important thread from Chapter 3 reappears in the NLI work: a preference for binary representations when using lexical features. The third is more implicit but has been suggested both in our discussion of prescriptivism in Chapter 2 and the discussion of non-native language learner assistance in Chapter 4: helping

non-native speakers improve their style requires, I believe, both an understanding of stylistic variation in the target language as well as a sense of how the L1 is influencing the style of the learner. I have not created a system for lexical style error correction here, but it is one application where the diverse aspects of style that I have addressed here would all be relevant, and one where the lexicon has not been a major focus as of yet, despite the obvious challenges that non-native learners face in this regard.

Long-term research goals aside, there are plenty of avenues for further research in native language identification: there are more and more resources becoming available for this task, and our research into corpus variables (Section 5.6) has far from settled the question of how we should (or shouldn't) use them. In our contribution to the NLI shared task, we effectively used a simple metric to carry out training feature selection, an idea which could be merged with our detailed analysis of the effects of variables and metrics which represent these variables in Section 5.6. Although Tree Substitution Grammar fragments were not effective in the shared task, we believe, as we do with our expansion of general stylistic resources (i.e. Chapter 3), that in the long run there will be benefit in taking a fully lexicogrammatical approach, perhaps in combination with automated error correction to improve extractions of these patterns. And, although this takes us beyond the lexicon, it would also be interesting to look at 'stylistic' preferences at the level of the discourse, that is, how L1 backgrounds affect the organization of ideas.

The lexical focus of this work, then, is not to suggest a move to the other extreme, i.e. that the lexicon is the *only* thing relevant to style. Certainly, other features have had a role throughout our discussion here; for example, when we looked at building readability lexicons we found that simple textual metrics, averaged across many instances, were a good proxy for lexical co-occurrence. But I think, based on what we have seen here, there is little doubt that the role of the lexicon in stylistic variation has been underappreciated. Our work in lexicon

induction shows, I hope, how vast this space really is (in terms of the lexical items involved), and our application of these lexicons to various tasks shows that this information could be useful, while suggesting other hurdles to be overcome. What is especially promising about our lexical focus to style is the way that improvements in statistical modeling can be applied to extract more and more information from larger and larger corpora; the sheer size of the lexicon, and the breadth of stylistic phenomena, offer a wealth of long-term potential for continued, robust improvement in relevant tasks.

References

Abrams, M. H. 1999. *A Glossary of Literary Terms*. Harcourt Brace, Toronto, 7th edition.

Alpaydin, Ethem. 2010. *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, August.

Argamon, Shlomo and Mosche Koppel. 2010. The rest of the story: Finding meaning in stylistic variation. In Argamon, Shlomo, Kevin Burns, and Shlomo Dubnov, editors, *The Structure of Style*, pages 79–112. Springer.

Argamon, Shlomo, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Information Systems*.

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 7:91–109.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Attali, Yigal and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Attali, Yigal and Don Powers. 2007. A developmental writing scale. Technical report, Educational Testing Service.

Attali, Yigal. 2007. Construct validity of *e-rater* in scoring TOEFL essays. Technical report, Educational Testing Service.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May.

Bagga, Amit and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING '98)*, pages 79–85, Montreal, Quebec, Canada.

Baljko, Melanie and Graeme Hirst. 1999. The importance of subjectivity in computational stylistic assessment. *Text Technology*, 9(1):5–17.

Baroni, Marco and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Bautin, Mikhail, Lohit Vijayaren, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM '08)*.

Bedient, Calvin. 1986. *He Do the Police in Different Voices: The Waste Land and its protagonist*. University of Chicago Press, Chicago.

Beeferman, Doug, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP '97)*, pages 35–46.

Bestgen, Yves, Sylviane Granger, and Jennifer Thewissen. 2012. Error patterns and automatic identification. In Jarvis, Scott and Scott A. Crossley, editors, *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, pages 127–153. Multilingual Matters.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. John Benjamins Publishing Company.

Birch, David. 1995. Introduction. In Leckie-Tarry, Helen, editor, *Language and Context: A Functional Linguistic Theory of Register*. Pinter.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. Technical report, Educational Testing Service.

Blei, David M. and John D. Lafferty. 2007. Correlated topic models. *Annals of Applied Statistics*, 1(1):17–35.

Blei, David M. and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '01, pages 343–348.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brants, Thorsten and Alex Franz. 2006. *Web 1T 5-gram Corpus Version 1.1*. Google Inc.

Brooke, Julian and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. Presented at the 2011 Conference of Learner Corpus Research (LCR2011).

Brooke, Julian and Graeme Hirst. 2012a. Facets of formality: A dimension of register in a sociolinguistic corpus. In *Georgetown University Round Table on Languages and Linguistics 2012 (GURT 2012)*.

Brooke, Julian and Graeme Hirst. 2012b. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, pages 779–784, Istanbul, Turkey.

Brooke, Julian and Graeme Hirst. 2012c. Paragraph clustering for intrinsic plagiarism detection using a stylistic vector-space model with extrinsic features. In *Notebook for PAN 2012 Lab at CLEF 12*, Rome.

Brooke, Julian and Graeme Hirst. 2012d. Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

Brooke, Julian and Graeme Hirst. 2013a. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 13th Annual Conference of the North American Chapter of the Association for Computational Lingusitics*.

Brooke, Julian and Graeme Hirst. 2013b. Multidimensional analysis vs. latent semantic analysis for constructing a register space: Are hand-coded features needed, or is bag-of-words enough? Presented at the International Conference on Genre- and Register-related Text and Discourse Features in Multilingual Corpora (LSB 13).

Brooke, Julian and Graeme Hirst. 2013c. Using other corpora in the 2013 NLI shared task. In *Proceedings of the 8th Workshop on Building Educational Applications Using NLP*, Atlanta.

Brooke, Julian and Sali Tagliamonte. 2012. Hunting the linguistic variable: Using computational techniques for data exploration and analysis. In *Georgetown University Round Table on Languages and Linguistics 2012 (GURT 2012)*.

Brooke, Julian, Tong Wang, and Graeme Hirst. 2010a. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Beijing.

Brooke, Julian, Tong Wang, and Graeme Hirst. 2010b. Inducing lexicons of formality from corpora. In *Proceedings of the Language Resources and Evaluation Conference (LREC '10), Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, Malta.

Brooke, Julian, Tong Wang, and Graeme Hirst. 2011. Clipping prediction with latent semantic analysis. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP '11)*.

Brooke, Julian, Adam Hammond, and Graeme Hirst. 2012a. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In *Proceedings of the 1st Workshop on Computational Literature for Literature (CLFL '12)*, Montreal.

Brooke, Julian, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst. 2012b. Building readability lexicons with unannotated corpora. In *Proceedings of the 1st Workshop on Predicting and Improving Text Readability for target reader populations*.

Brooke, Julian, Graeme Hirst, and Adam Hammond. 2013. Clustering voices in *The Waste Land*. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL '13)*, Atlanta.

Bruzzone, Lorenzo and Mattia Marconcini. 2010. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:770–787.

Burnard, Lou. 2000. User reference guide for British National Corpus. Technical report, Oxford University.

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98), Workshop on Discourse Relations and Discourse Marking*, Montreal, Canada.

Burstein, Jill, Jane Shore, John Sabatini, Yong-Won Lee, and Matthew Ventura. 2007. The automated text adaptation tool. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '07), Software Demonstrations*, pages 3–4, Rochester, New York.

Burstein, Jill. 2003. The e-rater scoring engine: Automated Essay Scoring with natural language processing. In Shermis, Mark D. and Jill Burstein, editors, *Automated Essay Scoring: A Cross Disciplinary Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Burton, Kevin, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA.

Buscail, Laurie and Patrick Saint-Dizier. 2009. Textual and stylistic error detection and correction: Categorization, annotation and correction strategies. Paper presented at the International Symposium on Natural Language Processing (IEEE-SNLP 2009).

Bykh, Serhiy and Detmar Meurers. 2012. Native language identification using recurring *n*-grams – investigating abstraction and domain dependence. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

Campbell, Sarah and Celia Roberts. 2007. Migration, ethnicity and competing discourses in the job interview: synthesizing the institutional and personal. *Discourse and Society*, 18(3):243–271.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying English text for language impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 269–270, Bergen,Norway.

Carter, Ronald. 1998. *Vocabulary: Applied Linguistic Perspectives*. Routledge, London.

Catt, Mark and Graeme Hirst. 1990. An intelligent CALI system for grammatical error diagnosis. *Computer Assisted Language Learning*, 3:3–26.

Chambers, J.K. 2006. Canadian raising retrospect and prospect. *Canadian Journal of Linguistics*, 51(2):105–118.

Chang, Pi-Chuan, Michel Gally, and Christopher Manning. 2008a. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the ACL '08 Third Workshop on Statistical Machine Translation*.

Chang, Yu-Chia, Jason S. Chang, Hao-Jan Chen, and Hsien-Chin Liou. 2008b. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Neural Information Processing Systems (NIPS '09)*.

Changizi, Mark A. 2008. Economically organized hierarchies in WordNet and the Oxford English Dictionary. *Cognative Systems. Research*, 9(3):214–228, June.

Cheville, Julie. 2004. Automated scoring technologies and the rising influence of error. *The English Journal*, 93(4):47–52.

Chodorow, Martin and Jill Burstein. 2004. Beyond essay length: Evaluating e-rater's performance on TOEFL essays. Technical report, Educational Testing Service.

Chung, Gregory K.W.K. and Eva L. Baker. 2003. Issues in the reliability and validity of automated scoring of constructed responses. In Shermis, Mark D. and Jill Burstein, editors, *Automated Essay Scoring: A Cross Disciplinary Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Coleman, Meri and T.L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Collins-Thompson, Kevyn and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science Technology*, 56(13):1448–1462.

Coltheart, Max. 1980. *MRC Psycholinguistic Database User Manual: Version 1*. Birkbeck College.

Comrie, Bernard, editor. 1987. *The World's Major Languages*. Oxford University Press, Oxford.

Cook, Paul and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the NAACL HLT 2009 Workshop on Computational Approaches to Linguistic Creativity*, pages 71–79, Boulder, Colorado.

Cooper, John Xiros. 1987. *T.S. Eliot and the politics of voice: The argument of The Waste Land*. UMI Research Press, Ann Arbor, Mich.

Cox, David R. and Peter A.W. Lewis. 1966. *The Statistical Analysis of Series of Events*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Crammer, Koby and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292.

Crossley, Scott A. and Danielle S. McNamara. 2012. Detecting the first language of second language writers using automated indicies of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In Jarvis, Scott and Scott A. Crossley, editors, *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*. Multilingual Matters.

Crystal, David and Derek Davy. 1969. *Investigating English Style*. Indiana University Press.

Dahlmeier, Daniel and Hwee Tou Ng. 2011. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 107–117, Edinburgh, Scotland, UK.

Dale, Edgar and Jeanne Chall. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA.

Daumé, Hal and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genova, Italy.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

DiMarco, Chrysanne and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–499.

Dolch, Edward William. 1948. *Problems in Reading*. The Garrard Press.

Doyle, Gabriel and Charles Elkan. 2009. Accounting for burstiness in topic models. In *International Conference on Machine Learning (ICML '09)*.

Duggan, Joseph J. 1973. *The Song of Roland: Formulaic style and poetic craft*. University of California Press.

Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 148–155, Bethesda, Maryland, United States. ACM.

Eisenstein, Jacob and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08*, EMNLP '08, pages 334–343.

Emigh, William and Susan C. Herring. 2005. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*, Washington, DC.

Ericsson, Patricia and Richard Haswell, editors. 2006. *Machine scoring of student essays: truth and consequences*. Utah State University Press, Logan.

Erk, Katrin and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906, Honolulu, Hawaii.

Estival, Dominique, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING '07)*, pages 263–272.

Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation(LREC)*, Genova, Italy.

Fabrigar, Leandre R., Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Faruqui, Manaal and Sebastian Padó. 2011. "I thou thee, thou traitor": Predicting formal vs. informal address in English literature. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Fette, Ian, Norman Sadeh, and Anthony Tomasic. 2007. Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pages 649–656, Banff, Alberta, Canada.

Finn, Aidan, Nicholas Kushmerick, and Barry Smyth. 2002. Genre classification and domain transfer for information filtering. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, pages 353–362.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Follett, Wilson. 1966. *Modern American Usage*. Hill & Wang, New York.

Fowler, H. W. and F. G. Fowler. 1906. *The King's English*. Clarendon Press, Oxford, 2nd edition.

Fowler, H. W. 1968. *A Dictionary of Modern English Usage*. Oxford University Press, 2nd edition.

Francis, Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.

Futagi, Yoko, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21:353–367.

Galley, Michel, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL '03)*, ACL '03, pages 562–569, Sapporo, Japan.

Garera, Nikesh and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09)*, pages 710–718, Singapore.

Garner, Bryan. 2009. *Garner's Modern American Usage*. Oxford University Press, 3rd edition.

Glover, Angela and Graeme Hirst. 1996. Detecting stylistic inconsistencies in collaborative writing. In Sharples, Mike and Thea van der Geest, editors, *The New Writing Environment: Writers at Work in a World of Technology*. Springer-Verlag, London, UK.

Godfrey, J.J., E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520.

Golcher, Felix and Marc Reznicek. 2011. Stylometry and the interplay of title and L1 in the different annotation layers in the Falko corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*, Berlin.

Golub, Gene H. and Charles F. Van Loan. 1996. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA.

Graf, Arnulf B. A. and Silvio Borer. 2001. Normalization in support vector machines. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, pages 277–282.

Graham, Neil, Graeme Hirst, and Bhaskara Marthi. 2005. Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4):397–415.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Gregory, Michael and Susanne Carroll. 1978. *Language and Situation: Language Varieties and their Social Contexts*. Routledge and Kegan Paul, Boston.

Gunning, Robert. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

Guthrie, David. 2008. *Unsupervised Detection of Anomalous Text*. Ph.D. thesis, University of Sheffield.

Halliday, M.A.K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Halliday, M.A.K. 1994. *Introduction to Functional Grammar*. Edward Arnold, London, 2nd edition.

Hammond, Adam, Julian Brooke, and Graeme Hirst. 2013. A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL '13)*, Atlanta.

Hammond, Adam. 2013. He do the police in different voices: Looking for voices in *The Waste Land*. Seminar: "Mapping the Fictional Voice" American Comparative Literature Association (ACLA).

Han, Bo, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

Hassan, Ahmed and Dragomir Radev. 2010. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (ACL '10), pages 395–403.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL '97)*, ACL '98, pages 174–181, Madrid, Spain.

Hayakawa, S.I., editor. 1994. *Choose the Right Word*. HarperCollins Publishers, second edition. Revised by Eugene Ehrlich.

Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, ACL '94, pages 9–16.

Hearst, Marti A. 1998. Automated discovery of WordNet relations. In Fellbaum, Christiane, editor, *WordNet: An Electronic Lexical Database*, pages 131–153. MIT Press.

Heylighen, Francis and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340.

Hoffman, Matthew D., David M. Blei, and Francis R. Bach. 2010. Online learning for latent Dirichlet allocation. In *Neural Information Processing Systems (NIPS '10)*, pages 856–864.

Hovy, Eduard H. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2):153–197.

Hsu, Chih-Wei and Chih-Jen Lin. 2002. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425.

Inkpen, Diana and Graeme Hirst. 2006. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262.

Inkpen, Diana. 2007. A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing*, 4(1):1–17.

Ishikawa, Shin'ichiro. 2011. A new horizon in learner corpus studies: The aim of the ICNALE project. In *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11. University of Strathclyde Press, Glasgow, UK.

Ito, Rika and Sali A. Tagliamonte. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, 32(2):257–279.

Jarvis, Scott and Magali Paquot. 2012. Exploring the role of *n*-grams in L1 identification. In Jarvis, Scott and Scott A. Crossley, editors, *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, pages 71–105. Multilingual Matters.

Johnson-Laird, Philip. 1983. *Mental Models: Towards a Cognitive Science Theory of Language, Inference and Consciousness*. Cambridge University Press, New York.

Jones, Karen Sparck. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–22.

Joos, Martin. 1961. *The Five Clocks*. Harcourt, Brace and World, New York.

Kaji, Nobuhiro and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*.

Kamps, Jaap, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*.

Kane, Thomas S. 1983. *The Oxford Guide to Writing*. Oxford Univeristy Press.

Kang, Feng, Rong Jin, and Rahul Sukthankar. 2006. Correlated label propagation with application to multi-label learning. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*.

Karlgren, Jussi and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Kyoto, Japan.

Keith, Timothy. 2003. Validity of automated essay scoring systems. In Shermis, Mark D. and Jill Burstein, editors, *Automated Essay Scoring: A Cross Disciplinary Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, pages 32–38, Madrid, Spain.

Kestemont, Mike, Kim Luyckx, and Walter Daelemans. 2011. Intrinsic plagiarism detection using character trigram distance scores. In *Proceedings of the PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse*.

Kidwell, Paul, Guy Lebanon, and Kevyn Collins-Thompson. 2009. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 900–909, Singapore.

Kincaid, J. Peter, Robert. P. Fishburne Jr., Richard L. Rogers, and Brad. S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Kochmar, Ekaterina. 2011. Identification of a writer's native language by error analysis. Master's thesis, University of Cambridge.

Koppel, Moshe and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Koppel, Moshe, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pages 624–628, Chicago, Illinois, USA.

Koppel, Moshe, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Kreidler, Charles. 1979. Creating new words by shortening. *Journal of English Linguistics*, 13(24):24–36.

Kullback, Solomon and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Labov, William. 1972. *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.

Lahiri, Shibamouli, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 446–457.

Landauer, Thomas K. and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Langkilde-Geary, Irene. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 2nd International Conference on Natural Language Generation*, New York.

Lanham, Richard A. 1974. *Style:An Anti-textbook*. Paul Dry Books.

Leacock, Claudia, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool.

Leckie-Tarry, Helen. 1995. *Language and Context: A Functional Linguistic Theory of Register*. Pinter.

Lee, Yong-Won, Claudia Gentile, and Robert Kantor. 2010. Toward Automated Multi-trait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics*, 31(3):391–417.

Li, Hanhong and Alex C. Feng. 2011. Age tagging and word frequency for learners' dictionaries. In Newman, John, Harald Baayen, and Sally Rice, editors, *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Rodopi, Amsterdam.

Lin, Dekang and Patrick Pantel. 2001. DIRT - discovery of inference rules from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328.

Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 42(2):67–73.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Malioutov, Igor and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL '06)*, pages 25–32, Sydney, Australia.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Max, Aurélien. 2006. Writing for language-impaired readers. In *Proceedings of 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing '06)*, pages 567–570.

McKenna, C. W. F. and A. Antonia. 2001. The statistical analysis of style: Reflections on form, meaning, and ideology in the 'Nausicaa' episode of *Ulysses*. *Literary and Linguistic Computing*, 16(4):353–373.

Michos, S. E., N. Fakotakis, and G. Kokkinakis. 1999. Enhancing text retrieval by using advanced stylistic techniques. *Journal of Intelligent Robotics Systems*, 26(2):137–156.

Milroy, James and Leslie Milroy. 1999. *Authority in Language: Investigating Standard English*. Routledge & Kegan Paul, New York, 3rd edition.

Mohammad, Saif and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles.

Mohammad, Saif, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, EMNLP '09, pages 599–608, Singapore.

Morik, Katharina, Peter Brockhausen, and Thorsten Joachims. 1999. Combining statistical learning with a knowledge-based approach — a case study in intensive care monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 268–277.

Mosquera, Alejandro and Paloma Moreda. 2012. A qualitative analysis of informality levels in web 2.0 texts: The facebook case study. In *Proceedings of the LREC workshop:@ NLP can u tag# user_generated_content*, pages 23–29.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '10)*, pages 100–108, Los Angeles, California.

Newman, David, Edwin V. Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS '11)*.

Oberreuter, Gabriel, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. 2011. Approaches for intrinsic and external plagiarism detection. In *Proceedings of the PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse*.

Odlin, Terence. 1989. *Language Transfer*. Cambridge University Press.

Okazaki, Naoaki and Sophia Ananiadou. 2006. A term recognition approach to acronym recognition. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL '06)*.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319.

Paiva, Daniel S. and Roger Evans. 2004. A framework for stylistically controlled generation. In *Proceedings of the 3rd International Conference on Natural Language Generation*, New Forest, UK.

Paiva, Daniel S. and Roger Evans. 2005. Empirically-based control of natural language generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 58–65, Ann Arbor, Michigan.

Pan, Sinno Jialin and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10).

Paolillo, John C. 2000. Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics*, 36(2):215–259.

Payette, Julie and Graeme Hirst. 1992. An intelligent computer assistant for stylistic instruction. *Computers and the Humanities*, 26(2):87–102.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Erlbaum Publishers, Mahwah, NJ.

Petersen, Sarah E. and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1):89–106.

Peterson, Kelly, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Pevzner, Lev and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, March.

Plank, Barbara and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA, June.

Ponisciak, Steve and Valen Johnson. 2003. Bayesian analysis of essay grading. In Shermis, Mark D. and Jill Burstein, editors, *Automated Essay Scoring: A Cross Disciplinary Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Porteous, Ian, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, pages 569–577, Las Vegas, Nevada, USA.

Pullum, Geoffrey K. 2009. 50 years of stupid grammar advice. *The Chronicle of Higher Education*, 55(32):B15.

Quinlan, Thomas, Derrick Higgins, and Susanne Wolff. 2009. Evaluating the construct-coverage of the e-rater scoring engines. Technical report, Educational Testing Service.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman Group.

Ramos, Margarita Alonso, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez, and Sabela Prieto González. 2010. Towards a motivated annotation schema of collocation errors in learner corpora. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*.

Rao, Delip and Deepak Ravichandra. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Lingusitics*, Athens, Greece.

Reisinger, J., A. Waters, B. Silverthorn, and R. Mooney. 2010. Spherical topic models. In *International Conference on Machine Learning (ICML '10)*.

Rosenthal, Sara and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Rozovskaya, Alla and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Scharber, Cassandra, Sara Dexter, and Eric Riedel. 2008. Students experiences with an automated essay scorer. *Journal of Technology, Learning, and Assessment*, 7(1).

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March.

Selinker, Larry. 1992. *Rediscovering Interlanguage*. Longman, London.

Sheika, Fadi Abu and Diana Inkpen. 2012. Learning to classify documents according to formal and informal style. *Linguistic Issues in Language Technology*, 8.

Shermis, Mark D. and Jill Burstein, editors. 2003. *Automated Essay Scoring: A Cross-Disciplinary Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Si, Luo and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*, pages 574–576, Atlanta, Georgia, USA.

Simonton, Dean Keith. 1990. Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. *Computers and the Humanities*, 24(4):251–264.

Stamatatos, Efstathios. 2009a. Intrinsic plagiarism detection using character *n*-gram profiles. In *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and, Social Software Misuse (PAN-09)*, pages 38–46. CEUR Workshop Proceedings, volume 502.

Stamatatos, Efstathios. 2009b. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Stein, Benno, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilivie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Strunk, William and E.B. White. 1979. *The Elements of Style*. Macmillan, 3rd edition.

Swanson, Ben and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL '12)*, pages 193–197, Jeju, Korea.

Swanson, Ben and Eugene Charniak. 2013. Extracting the native language signal for second language acquisition. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '13)*.

Taboada, Maite, Caroline Anthony, and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC '06)*, pages 427–432.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Tagliamonte, Sali A. 2005. So who? Like how? Just what? Discourse markers in the conversations of English speaking youth. *Journal of Pragmatics*, 37(11):1896–1915.

Tagliamonte, Sali A. 2006. "So cool, right?": Canadian English entering the 21st century. *Canadian Journal of Linguistics*, 51(2):309–331.

Tagliamonte, Sali A. 2011. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Wiley-Blackwell Publishers.

Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 133–140, Ann Arbor, Michigan.

Tanaka-Ishii, Kumiko, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

Tetreault, Joel, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

Tetreault, Joel, Daniel Blanchard, and Aoife Cahill. 2013. Summary report on the first shared task on native language identification. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*, Atlanta, GA, USA, June. Association for Computational Linguistics.

Tomokiyo, Laura Mayfield and Rosie Jones. 2001. You're not from 'round here, are you?: naïve Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01)*, pages 1–8, Pittsburgh, Pennsylvania.

Tono, Yukio, Yuji Kawaguchi, and Makoto Minegishi, editors. 2012. *Developmental and Cross-linguistic Perspectives in Learner Corpus Research*. John Benjamins, Amsterdam/Philadelphia.

Touch Press LLP. 2011. *The Waste Land* app. http://itunes.apple.com/ca/app/the-waste-land/id427434046?mt=8 .

Trudgill, Peter. 2000. *Sociolinguistics: An Introduction to Language and Society (4th Edition)*. Penguin Books, London.

Tsur, Oren and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA '07)*, pages 9–16, Prague, Czech Republic.

Turney, Peter and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141.

Turney, Peter D., Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Emperical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, United Kingdom.

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, Pennsylvania.

University of Chicago. 2003. *The Chicago Manual of Style*. University of Chicago Press.

Utiyama, Masao and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL '01)*, pages 499–506, Toulouse, France.

Vajjala, Sowmya and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 163–173, Montreal, Canada.

van Dijk, Teun A. 2008. *Discourse and Context: A Sociocognitive Approach*. Cambridge University Press.

van Halteren, Hans. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, pages 937–944, Manchester, UK.

van Oosten, Philip, Dries Tanghe, and Veronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, Malta.

Velikovich, Leonid, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 777–785, Los Angeles, California.

Wang, Tong and Graeme Hirst. 2010. Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, pages 1182–1190, Beijing.

Warshauer, Mark and Paige Ware. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2):1–24.

Williams, Joseph M. 1990. *Style: Towards Clarity and Grace*. University of Chicago Press.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Wong, Sze-Meng Jojo and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61.

Wong, Sze-Meng Jojo and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1600–1610, Edinburgh, Scotland, UK.

Wong, Sze-Meng Jojo, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, Jeju, Korea.

Woolf, Virginia. 1927. *To the Lighthouse*. Hogarth, London.

Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 180–189, Portland, Oregon.

Yuret, Deniz and Mehmet Ali Yatbaz. 2010. The noisy channel model for unsupervised word sense disambiguation. *Computational Linguistics*, 36(1):111–127, March.