AFRL-RH-WP-TR-2010-0038

Detecting the Difficulty Level of Foreign Language Texts

Raymond E. Slyh
Eric G. Hansen

Anticipate & Influence Behavior Division
Sensemaking & Organizational Effectiveness Branch

February 2010

Final Report for October 2005 to September 2009

Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Anticipate & Influence Behavior Division
Sensemaking & Organizational Effectiveness Branch
Wright-Patterson AFB OH 45433-7022

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2010-0038 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//	//SIGNED//
RAYMOND E. SLYH	GLENN W. HARSHBERGER
Work Unit Manager	Anticipate & Influence Behavior Division
Sensemaking & Organizational	Human Effectiveness Directorate
Effectiveness Branch	711th Human Performance Wing
	Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* February 2010	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* October 2005 – September 2009

4. TITLE AND SUBTITLE Detecting the Difficulty Level of Foreign Language Texts	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 62202F
6. AUTHOR(S) Raymond E. Slyh, Eric G. Hansen	**5d. PROJECT NUMBER** 7184
	5e. TASK NUMBER X0
	5f. WORK UNIT NUMBER 7184X07C
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Anticipate & Influence Behavior Division Sensemaking & Organization Effectiveness Branch	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711th Human Performance Wing Human Effectiveness Directorate Anticipate & Influence Behavior Division Sensemaking & Organizational Effectiveness Branch Wright-Patterson AFB OH 45433-7022	**10. SPONSOR/MONITOR'S ACRONYM(S)** 711 HPW/RHXS
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RH-WP-TR-2010-0038

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

88ABW cleared on 18Mar10, 88ABW-2010-1309.

14. ABSTRACT

This report describes experiments conducted on automatically determining the difficulty level of foreign language materials for the purpose of aiding teachers, students, and DoD linguists in finding suitable materials for supporting language learning and sustainment. The measure used as the indicator of difficulty is based on the Interagency Language Roundtable (ILR) proficiency scale, which is used to measure the proficiency levels of DoD linguists in listening, reading, speaking, writing, translating, and interpreting. The experiments described were conducted with a corpus of authentic Arabic and Mandarin Chinese materials from several genres that were hand-labeled for ILR level. The corpus contained materials at the 2, 2+, and 3 levels. ILR level detectors were built for these levels for both the original Arabic and Mandarin sources as well as for human-produced English translations of these sources. The detectors were based on statistical language modeling techniques. The equal error rates (EERs) obtained ranged from 12.4–49.4% depending on the language, ILR level, language model order, and various other factors related to the experimental design. In general, the performance was best for discriminating level 3 materials from level 2 and 2+ materials, with EERs ranging from 12.4–33.3% across the languages (and translations), language model level, and experimental design. The performance was worst for discriminating level 2+ materials from level 2 and 3 materials, with EERs ranging from 31.2–49.4%.

15. SUBJECT TERMS Interagency Language Roundtable, Linguist Proficiency, Readability, Text Difficulty, Text Classification, Arabic, Mandarin, Chinese

16. SECURITY CLASSIFICATION OF: Unclassified			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Raymond E. Slyh
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	42	**19b. TELEPHONE NUMBER** *(include area code)* NA

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This report describes experiments conducted on automatically determining the difficulty level of foreign language materials for the purpose of aiding teachers, students, and DoD linguists in finding suitable materials for supporting language learning and sustainment. The measure used as the indicator of difficulty is based on the Interagency Language Roundtable (ILR) proficiency scale, which is used to measure the proficiency levels of DoD linguists in listening, reading, speaking, writing, translating, and interpreting. The experiments described were conducted with a corpus of authentic Arabic and Mandarin Chinese materials from several genres that were hand-labeled for ILR level. The corpus contained materials at the 2, 2+, and 3 levels. ILR level detectors were built for these levels for both the original Arabic and Mandarin sources as well as for human-produced English translations of these sources. The detectors were based on statistical language modeling techniques. The equal error rates (EERs) obtained ranged from 12.4–49.4% depending on the language, ILR level, language model order, and various other factors related to the experimental design. In general, the performance was best for discriminating level 3 materials from level 2 and 2+ materials, with EERs ranging from 12.4–33.3% across the languages (and translations), language model level, and experimental design. The performance was worst for discriminating level 2+ materials from level 2 and 3 materials, with EERs ranging from 31.2–49.4%. Finally, the report makes a number of suggestions for future work.

1.0 INTRODUCTION

Air Force and Department of Defense (DoD) personnel are called upon to operate all over the world with the Global War on Terror, various coalition military operations, humanitarian relief operations, and other interactions with multinational partners. With its global reach and responsibilities, the DoD needs to monitor and understand ongoing situations, to anticipate new situations that will require responses, and to influence outcomes. Much of the information needed to effectively understand, anticipate, and influence these situations and to operate in them is found in foreign language speech and text. For this reason, military linguists are critical assets, and it is important that they sustain their language skills and even increase their proficiency levels.

Whether for sustainment or for increasing skill levels, it is important that linguists continually work with authentic foreign language materials of sufficient difficulty. Authentic materials are materials that have been written or spoken by native speakers of a language and that are intended for native speakers of the language. However, with numerous radio and television broadcasts and rapidly growing amounts of foreign language text available on the web, it is a difficult task for teachers, students, and DoD linguists to efficiently locate materials of the appropriate difficulty to support language learning and sustainment.

This report describes experiments conducted on automatically determining the difficulty level of foreign language materials for the purpose of aiding teachers, students, and DoD linguists in finding suitable materials for supporting language learning and sustainment. The measure used as the indicator of difficulty is based on the Interagency Language Roundtable (ILR) proficiency scale, which is used to measure the proficiency levels of DoD linguists in listening, reading, speaking, writing, translation, and interpretation (Interagency Language Roundtable, 2010). The experiments described were conducted with a corpus of authentic Arabic and Mandarin Chinese materials from several genres that were hand-labeled for ILR level. The corpus contained materials at the 2, 2+, and 3 levels. ILR level detectors were built for these levels for both the original Arabic and Mandarin sources as well as for human-produced English translations of these sources. The detectors were based on statistical language modeling techniques. The equal error rates (EERs) obtained ranged from 12.4–49.4% depending on the language, ILR level, language model order, and various other factors related to the experimental design. In general, the performance was best for discriminating level 3 materials from level 2 and 2+ materials, with EERs ranging from 12.4–33.3% across the languages (and translations), language model level, and experimental design. The performance was worst for discriminating level 2+ materials from level 2 and 3 materials, with EERs ranging from 31.2–49.4%.

An outline of this report is as follows. The next section describes the ILR proficiency scale descriptors, while Section 3.0 describes past work related to the work described here. Section 4.0 describes the corpus used for the experiments and the preprocessing applied to the data. Section 6.0 describes the process for building the detectors and the experimental results. Finally, Section 7.0 summarizes the conclusions and discusses the future work.

2.0 THE ILR PROFICIENCY SCALE

The ILR proficiency scale measures a person's proficiency in listening, reading, speaking, writing, translation, and interpretation (Interagency Language Roundtable, 2010). The scales for listening, reading, speaking, and writing consist of six "base levels" ranging from 0 (No Proficiency) to 5 (Functionally Native Proficiency), where each "base level" implies mastery of any previous "base level." In addition to the "base levels," there are "plus level" descriptors for levels 0 through 4 that are used to describe proficiency levels that substantially exceed one base level but do not fully meet the criteria for the next "base level." Thus, a linguist could be rated as R3/L2+/S1/W1+, which would indicate that he/she is at level 3 for reading, level 2+ for listening, level 1 for speaking, and level 1+ for writing. Linguists are assigned these skill levels in the various categories through authorized language examinations, and each base and plus level in a category contains a number of performance criteria that must be met to obtain that score. For example, the following are the level descriptors for 2, 2+, and 3 in reading from (Interagency Language Roundtable, 2010):

> **Reading Level 2 (Limited Working Proficiency)**
> Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform sorting and locating tasks with written texts that are well beyond their general proficiency level. The individual can read uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straightforward/high-frequency sentence patterns. The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. Is typically able to answer factual questions about authentic texts of the types described above.
>
> **Reading Level 2+ (Limited Working Proficiency, Plus)**
> Sufficient comprehension to understand most factual material in non-technical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance and/or intentionally disguised meaning.

> **Reading Level 3 (General Professional Proficiency)**
> Able to read within a normal range of speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and "read between the lines," (that is, understand the writers' implicit intents in text of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms.

It is important to note that, strictly speaking, the ILR scale is a rating of linguist proficiency and not of material difficulty. The descriptions for each base or plus level refer in many cases to skills that a linguist should posses, and of course, a piece of written text or audio doesn't posses any skills. It can't "comprehend" or "discern" anything. Despite this fact, the ILR scale can be used as a difficulty measure for texts and audio by assigning to the material the ILR proficiency level that a linguist generally would need in order to properly understand the material. Thus, a linguist rated at level 2 in reading generally should be able to fully understand texts rated at level 2, but would generally have trouble understanding at least some keys ideas in level 3 texts. For language learning and sustainment purposes, a linguist or student who wanted to increase his/her proficiency to level 3 would profit most by studying level 2+ and 3 materials. Materials below level 2 would likely be too easy and good only for occasional review, while materials at levels 4 and above would likely be too difficult.

3.0 RELATED WORK

There has been considerable past work related to various aspects of the work described here; however, the work on measures of readability would appear to be some of the most relevant. Researchers have been studying readability and devising various measures of readability since at least the 1920's (Kitson, 1921; Lively and Pressey, 1923; Vogel and Washburne, 1928). However, the Flesch Reading Ease (FRE) measure described in (Flesch, 1948) could well be considered to have touched off a flurry of activity that resulted in numerous classic measures such as the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), the Automated Readability Index (ARI) (Smith and Senter, 1967), the Coleman-Liau Index (CLI) (Coleman and Liau, 1975), the Gunning Fog Grade (FOG) (Gunning, 1952), the SMOG (Standard Measure of Gobbledygook) Grade (McLaughlin, 1969), the Spache Grade (SG) (Spache, 1953), and the Dale-Chall Readability Index (DCRI) (Dale and Chall, 1948; Chall and Dale, 1995). See (DuBay, 2004) for an extensive discussion and bibliography. The aforementioned measures were all developed for English; however, (Rabin, 1988) reports that similar readability formulas have also been developed for some foreign languages.

The FRE and the FKGL are both based on the average number of words per sentence, $\bar{W}$, and the average number of syllables per word, $\bar{S}$. The FRE is given as

$$\mathrm{FRE} = 206.835 - 1.015\bar{W} - 84.6\bar{S}.$$

Documents with FRE scores in the range of 90–100 are easily understandable by an average 11-year old student. Those with scores in the range of 60–70 are easily understandable by 13- to 15-year-old students, while documents with scores in the range of 0–30 are best understood by university graduates. The FKGL gives an estimate of the US grade level or years of formal education required to understand a text. The FKGL is given as

$$\mathrm{FKGL} = 0.39\bar{W} + 11.80\bar{S} - 15.59.$$

Users of the FRE and FKGL often found the process of counting syllables to be too time-consuming, so they devised similar measures that replaced $\bar{S}$ with the average number of characters per word, $\bar{C}$. The ARI is given as

$$\mathrm{ARI} = 4.71\bar{C} - 0.5\bar{W} - 21.43,$$

while the CLI is given as

$$\mathrm{CLI} = 5.89\bar{C} - \frac{29.5}{\bar{W}} - 15.8.$$

Both the ARI and the CLI produce an estimate of the US grade level needed to comprehend a given text.

The FOG and SMOG measures both depend on the concept of "complex" or "polysyllabic" words, and they both produce an estimate of the number of years of formal education required to understand a text. Let W_C be the number of complex words in a text, where a complex word is defined as a word with three or more syllables (excluding endings) which is not a name or a compound word. Let W be the number of total words in the text. The FOG is given as

$$\mathrm{FOG} = 0.4\bar{W} + 40\frac{W_C}{W}.$$

The SMOG grade is similar to the FOG in that it depends on the number of polysyllabic words, which are words of three or more syllables. Let $\bar{P}$ be the average number of polysyllabic words per sentence, then the SMOG grade is given as

$$\text{SMOG} = 1.043\sqrt{30\bar{P}} + 3.1291.$$

For the SMOG measure, it is recommended that one consider at least 30 sentences from the text, with at least ten from the start, ten from the middle, and ten from the end.

The next step in complexity among the classic readability measures includes the SG and the DCRI, which both depend on lists of familiar or everyday words. Words not on these lists are considered unfamiliar or difficult. Let W_U be the number of unfamiliar words in a text of length W words, then the SG is given as

$$\text{SG} = 0.141\bar{W} + 0.086\frac{W_U}{W} + 0.839,$$

while the DCRI is given as

$$\text{DCRI} = 0.0496\bar{W} + 15.79\frac{W_U}{W} + 3.6565.$$

The original DCRI used a list of 763 familiar words that fourth-grade students generally can understand. The revised list from (Chall and Dale, 1995) uses a list of 3,000 familiar words. There are also a number of guidelines about how to count hyphenated words, abbreviations, names of people and places, variants of the words on the list, *etc.* Finally, the raw DCRI can be converted to a corrected grade level as shown in Table 1.

Table 1: Conversion of Raw DCRI Scores to Corrected Grade Levels

Raw DCRI Score	Corrected Grade Levels
4.9 and below	4th Grade and below
5.0–5.9	5–6th Grade
6.0–6.9	7–8th Grade
7.0–7.9	9–10th Grade
8.0–8.9	11–12th Grade
9.0–9.9	13–15th Grade (College)
10.0 and above	16th–College Graduate

More recently, measures of readability based on statistical language modeling have been developed (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Heilman et al., 2007), and these can be seen as extending the basic ideas behind the SG and DCRI measures. In (Collins-Thompson and Callan, 2004), language models were built from English-language web documents that had been assigned one of twelve US school grade levels. It is important to note that the documents were "noisy" in that they contained navigation menus, links, *etc.* The language models consisted of smoothed unigram models using Good-Turing smoothing (Manning and Schütze, 1999) with additional smoothing across grade levels. The authors also applied their techniques to French documents. For both English and French documents, the authors were able to predict grade level with reasonable accuracy.

In (Schwarm and Ostendorf, 2005), the authors use support vector machines (Vapnik, 1995; Joachims, 1998, 1999) to classify the reading levels of English texts using the following features: 12 language model perplexities (Manning and Schütze, 1999), the average sentence length in words, the average number of syllables per word, the FKGL score, six out-of-vocabulary word rate scores, and features derived from parse trees (average parse tree height, average number of noun phrases, average number of verb phrases, and average number of SBARs[1]). The system was tested on documents assigned to US school grades 2–5, and it was shown that the system performed much better than the FKGL alone; however, it is unclear to what degree the various features contributed to the improved performance.

In (Heilman et al., 2007), the authors built on their previous work (Collins-Thompson and Callan, 2004) of using language models to predict reading level by also considering grammatical construction features extracted from parsing the texts. A set, G_1, of counts for 22 English grammatical constructions was extracted from sentence parse trees; the constructions included the use of passive voice, past participles, perfect verb tenses, relative clauses, continuous verb tenses, and modal verbs. The authors also considered a second set, G_2, of twelve grammatical construction features that could be extracted without parsing, including sentence length, various English verb forms (present, progressive, past, perfect, and continuous tenses), as well as part-of-speech labels for words. The authors used a k-Nearest Neighbor (kNN) classifier (Cover and Hart, 1967; Mitchell, 1997) for the grammatical features, and a final interpolated readability prediction, L_I, was given as

$$L_I = L_{LM} + C_{kNN} L_{GR},$$

where L_{LM} is the prediction of the language model system, L_{GR} is the prediction of one of the grammar-based classifiers, and C_{kNN} is a confidence value of the kNN prediction for the grammar-based classifier. Various system combinations were tested over two sets of materials. The first set of materials consisted of the English-language web documents considered in (Collins-Thompson and Callan, 2004). The second set consisted of textbook materials used to teach English as a second language; these materials were classified into four levels of 2 (beginning), 3, 4, and 5 (advanced). An important feature of the second set of materials is that the texts were scanned into electronic format, but they were not hand-corrected. Thus, both sets of texts contained noise, and it would be expected that the grammar-based classifiers would be more negatively affected by the noise than the language model classifier would be. For both sets of texts, the language model system substantially outperformed both the G_1 and G_2 systems when they were used alone. The interpolation of the language model system and either of the grammar-based systems outperformed the language model system, with the interpolation system using the G_1 features slightly outperforming the interpolation system using the G_2 features.

The work described in this report is most similar to that of (Collins-Thompson and Callan, 2004) in that we primarily consider the use of language modeling techniques to predict ILR level. However, unlike (Collins-Thompson and Callan, 2004), we consider bigram and trigram models in addition to unigram models. Further, we consider Arabic and Mandarin Chinese texts (along with their corresponding English translations), rather than French documents. Finally, (Collins-Thompson and Callan, 2004) did not consider any of the classic readability measures discussed in this section; however, in Section 5.0, we briefly examine the suitability of the classic FKGL, CLI, ARI, and FOG measures for predicting ILR level. We show that these measures are not well suited to predicting the ILR level for the data that we have.

[1] SBAR is defined as a clause introduced by a (possibly empty) subordinating conjunction.

4.0 THE GLOSS CORPUS AND PREPROCESSING

This section describes the corpus used for the ILR level detection experiments. It also describes the method used to partition the data into training and testing sets as well as the preprocessing that was applied to each file before feature extraction was performed.

4.1 Corpus Overview

The corpus used for the experiments consisted of material retrieved from the Global Language Online Support System (GLOSS) web site[2] managed by the Defense Language Institute Foreign Language Center. There were 194 Arabic and 361 Mandarin Chinese lessons retrieved from the site on 9 August 2007 and 19 February 2008, respectively. Each lesson contained a text passage in the native orthography (sometimes with a corresponding original audio track) and an English translation produced by a human translator. As discussed in Section 6.0, ILR level detection experiments were conducted on both the original source language and the English translations for each language.

Each lesson was rated for ILR level by trained human raters. In addition, each lesson was assigned to one of the following ten content/topical domains: culture (cul), economy (eco), environment (env), geography (geo), military (mil), politics (pol), science (sci), security (sec), society (soc), and technology (tec); however, there were not strict guidelines used to assign lessons to particular topics.[3]

The lessons in the corpus were not evenly distributed across ILR level or topic domain for either language as is illustrated in Tables 2–4 for Arabic and Tables 5–7 for Chinese. For both languages, the level 2 texts constitute almost 60% of the data. The level 2+ texts constitute almost 30% of the data, and the remainder (a little less than 10%) constitute the level 3 texts. For Arabic, the topics constituting at least 10% of the data across all the levels are society (28.4%), politics (18.6%), and economics (12.9%) for a total of 59.9%. For Chinese, the topics constituting at least 10% of the data across all the levels are society (22.2%), culture (18.3%), economics (11.6%), and environment (10.2%) for a total of 62.3%.

For a given language, there are a number of important differences in the percentage representation of various topics in the data between the three levels. In the Arabic data, economics texts constitute 13.0% and 16.7% of the files in levels 2 and 2+, respectively, but only 4.0% of the files in level 3. Geography texts constitute 10.4% of the files in level 2 but none of the files in levels 2+ and 3. Texts on politics are 15.7% of level 2 files and 20.4% of level 2+ but are 28.0% of level 3. Security texts constitute 13.0% of level 2 but approximately 4% for both levels 2+ and 3. Finally, the largest disparity in the Arabic data is seen in the texts on society, which constitute 42-44% of the data for levels 2+ and 3 but only 18.3% of the data for level 2. In the Chinese data, texts on culture constitute 17.9% of level 2, 21.9% of level 2+, but only 12.2% of level 3. Texts on the environment constitute 14.0% and 7.6% of levels 2 and 2+, respectively, but none of level 3. Political texts are 20.4% of level 3, but less than 10% of levels 2 and 2+. Finally, texts on society are 34.7% of level 3 but 21.0% or less of levels 2 and 2+.

4.2 Data Partitioning

The experiments discussed in this report used two different types of partitions of the data into training and testing sets. The experiments discussed in Subsections 6.1 and 6.2 used the

[2] `http://gloss.lingnet.org`

[3] T. Marius, Defense Language Institute Foreign Language Center, private communication, 8 October 2008.

Table 2: Distribution of Arabic Articles Across ILR Level and Topic

Level	cul	eco	env	geo	mil	pol	sci	sec	soc	tec	Total
2	9	15	6	12	7	18	6	15	21	6	115
2+	4	9	2	0	0	11	2	2	23	1	54
3	3	1	0	0	1	7	0	1	11	1	25
Total	16	25	8	12	8	36	8	18	55	8	194

Table 3: Percentage Distribution of Arabic Articles Across ILR Level and Topic

Level	cul	eco	env	geo	mil	pol	sci	sec	soc	tec	Total
2	4.6	7.7	3.1	6.2	3.6	9.3	3.1	7.7	10.8	3.1	59.3
2+	2.1	4.6	1.0	0.0	0.0	5.7	1.0	1.0	11.9	0.5	27.8
3	1.5	0.5	0.0	0.0	0.5	3.6	0.0	0.5	5.7	0.5	12.9
Total	8.2	12.9	4.1	6.2	4.1	18.6	4.1	9.3	28.4	4.1	100.0

Table 4: Percentage Distribution of Arabic Articles Across Topic Within ILR Level

Level	cul	eco	env	geo	mil	pol	sci	sec	soc	tec	Total
2	7.8	13.0	5.2	10.4	6.1	15.7	5.2	13.0	18.3	5.2	100.0
2+	7.4	16.7	3.7	0.0	0.0	20.4	3.7	3.7	42.6	1.9	100.0
3	12.0	4.0	0.0	0.0	4.0	28.0	0.0	4.0	44.0	4.0	100.0

Table 5: Distribution of Chinese Articles Across ILR Level and Topic

Level	cul	eco	env	geo	mil	pol	sci	sec	soc	tec	Total
2	37	26	29	12	12	14	12	11	41	13	207
2+	23	13	8	4	9	9	11	2	22	4	105
3	6	3	0	1	1	10	5	4	17	2	49
Total	66	42	37	17	22	33	28	17	80	19	361

Table 6: Percentage Distribution of Chinese Articles Across ILR Level and Topic

Level	cul	eco	env	geo	mil	pol	sci	sec	soc	tec	Total
2	10.2	7.2	8.0	3.3	3.3	3.9	3.3	3.0	11.4	3.6	57.3
2+	6.4	3.6	2.2	1.1	2.5	2.5	3.0	0.6	6.1	1.1	29.1
3	1.7	0.8	0.0	0.3	0.3	2.8	1.4	1.1	4.7	0.6	13.6
Total	18.3	11.6	10.2	4.7	6.1	9.1	7.8	4.7	22.2	5.3	100.0

Table 7: Percentage Distribution of Chinese Articles Across Topic Within ILR Level

Level	cul	eco	env	geo	mil	pol	sci	sec	soc	tec	Total
2	17.9	12.6	14.0	5.8	5.8	6.8	5.8	5.3	19.8	6.3	100.0
2+	21.9	12.4	7.6	3.8	8.6	8.6	10.5	1.9	21.0	3.8	100.0
3	12.2	6.1	0.0	2.0	2.0	20.4	10.2	8.2	34.7	4.1	100.0

following method for partitioning the data. For each language, the complete set of articles was partitioned into four subsets that were nearly equal in the number of documents. For each ILR level, the four subsets were balanced as much as possible for topic; however, as discussed in the prior subsection, there were clearly limits on this topic balancing. The subsets were then used in a round-robin fashion to generate multiple training and testing sets on which to run experiments. This was done to test the robustness of the algorithms and to generate more detection scores to smooth out the performance plots seen in Section 6.0. Table 8 shows the combination of the four subsets and how twelve experiments were conducted for each language and parameter setting. In all cases, two of the subsets were used for training, and the remaining two subsets were individually used for testing.

Table 8: Subset Combinations Used to Generate Experiment Sets for Subsections 6.1 and 6.2

Training Subsets	Testing Subset
s0 + s1	s2
s0 + s1	s3
s0 + s2	s1
s0 + s2	s3
s0 + s3	s1
s0 + s3	s2
s1 + s2	s0
s1 + s2	s3
s1 + s3	s0
s1 + s3	s2
s2 + s3	s0
s2 + s3	s1

For the experiments discussed in Subsection 6.3, training and testing data sets were created using the Fisher-Yates shuffle (Fisher and Yates, 1948; Durstenfeld, 1964; Knuth, 1998). For the set of articles of a given language and ILR level, the Fisher-Yates shuffle was used to create a random training and testing set so as to maintain a proportion of 80% training and 20% testing. This process was repeated to create 100 random training/testing sets for each language. These splits were not balanced for topic, so one would expect some degree of variability between the detection results for the various splits.

The number of articles and the number of tests for the two experimental setups (original round-robin splits and Fisher-Yates splits) are as follows:

Method 1—Round-Robin file lists:
Arabic: 4 splits (≈49 trials each), evaluated 3 times generated 582 trials
Chinese: 4 splits (≈90 trials each), evaluated 3 times generated 1083 trials

Method 2—Fisher-Yates file lists:
Arabic: 39 trials run 100 times generated 3900 trials
Chinese: 73 trials run 100 times generated 7300 trials

4.3 Preprocessing

Each text file was preprocessed to handle a number of issues related to punctuation, spurious characters, case, *etc.* The preprocessing steps were as follows:

- Each token in a file was converted to upper case if possible
- Any digit separators were removed, and decimal points were converted to the word "POINT" surrounded by spaces
- Any commas, braces, brackets, parentheses, slashes, colons, asterisks, various long or repeated dashes, and quotes were converted to spaces, except that apostrophes in contractions were retained
- The abbreviation "KM" was converted to "KILOMETERS"
- Any two- or three-letter acronyms with periods were normalized to their corresponding letters without the intervening periods.
- Each sentence (as determined by sentence-ending punctuation of periods, question marks, exclamation points, semicolons, or the foreign language equivalents of these punctuation marks) was augmented with sentence starting and ending tags, `<s>` and `<\s>`, respectively
- The ampersand sign, `&`, was converted to the word "AND" surrounded by spaces
- The dollar sign, $, followed by zero or more white space characters and one or more digits was converted to the digit sequence followed by the word "DOLLARS" with a space between the digit string and the word "DOLLARS"
- The percent sign, `%`, was converted to the word "PERCENT" with a leading space

4.4 Chinese Character Segmentation

Chinese is not normally segmented into words (in other words, the characters run together without intervening white space to denote word boundaries), and the original Chinese files followed this convention. For Chinese, two sets of ILR detection experiments were conducted on the Chinese source language files, one with all of the Chinese characters separated by white space (so that each character is a token) and one where the characters were automatically segmented into words using a Chinese word segmenter provided by the Linguistic Data Consortium (LDC). The word segmenter was `mansegment.perl` Version 1.0 written by Zhibiao Wu in 1999.[4]

[4] `http://projects.ldc.upenn.edu/Chinese/`

5.0 CLASSIC READABILITY MEASURES AND ILR

In this section, we briefly consider four of the classic readability measures discussed in Section 3.0—namely, the FKGL, CLI, ARI, and FOG measures—for their suitability in detecting ILR level. These measures were originally developed for English, and applying them to Arabic and Chinese would likely require extra research and/or development. To get a quick assessment of the suitability of these measures, we computed them on the English translations of the source language texts in our data set. The four readability measures were computed for each text using the `style` command available as part of the GNU `diction` software package.[5] For each of these measures, we fit Gaussian models to the scores by computing the mean and variance for each of the three ILR levels (2, 2+, and 3) and plotting the resulting Gaussian models.

Figure 1 shows the resulting Gaussian models plotted for each of the four measures when the models are built from the English translations of the Arabic texts. Figure 2 shows the corresponding plot when the models are built from the English translations of the Chinese texts. Regardless of the source language or readability measure, one can see a substantial overlap in the models for the three ILR levels. One possible explanation for the substantial overlap is that these four measures are not good discriminators of the ILR levels; however, a second possible explanation is that the English translations don't accurately reflect the same degree of difficulty as the original source language texts (at least in terms of features such as average sentence length, average number of syllables, or average number of characters). In the next section, we show that language modeling techniques can detect the ILR levels for the English documents to roughly the same degree as they can detect the ILR levels for the original source documents, so the remainder of this report focuses on the use of language modeling techniques.

[5] `http://www.gnu.org/software/diction/diction.html`

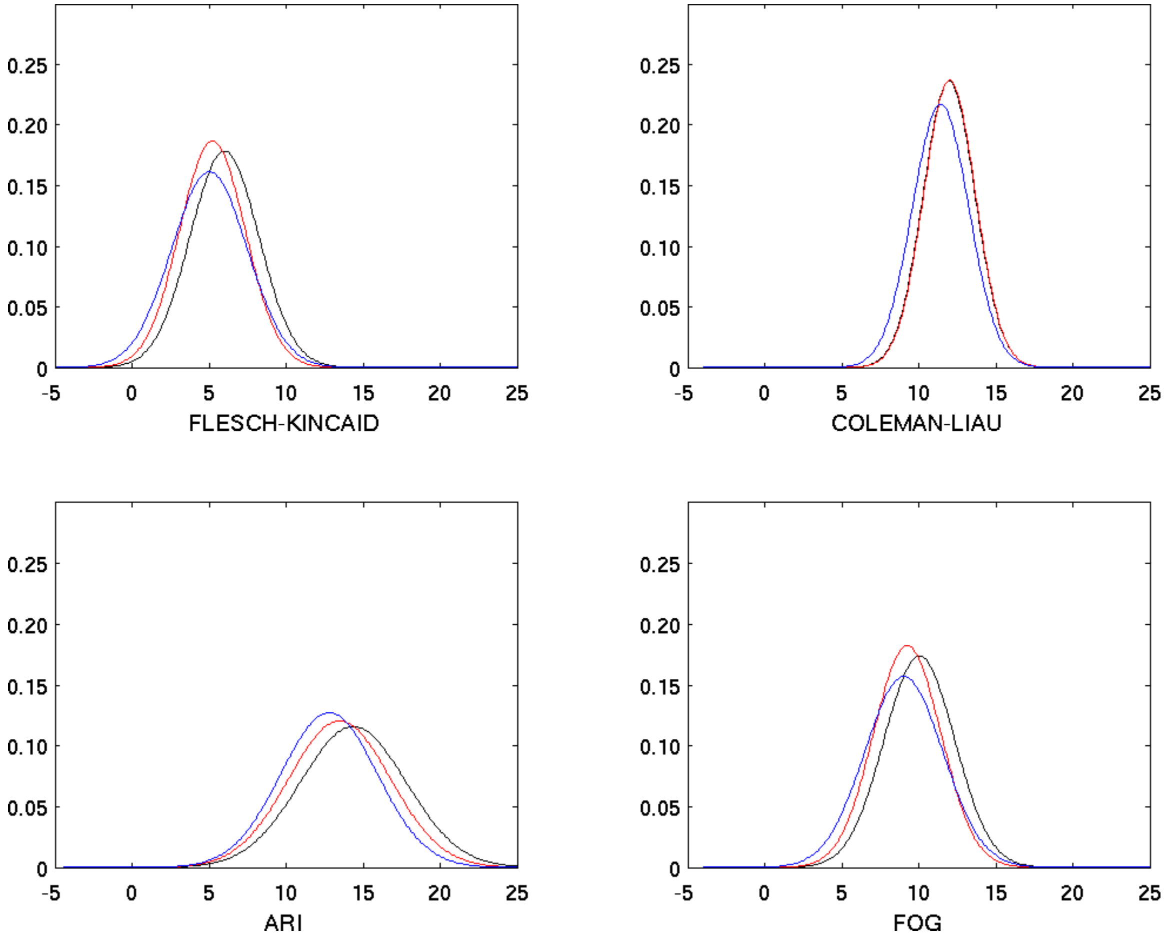

Figure 1: Gaussian Models of Flesch-Kincaid, Coleman-Liau, ARI, and FOG Measures Versus ILR Level for English Translations of Arabic Source Data (the results for levels 2, 2+, and 3 are in black, red, and blue, respectively)

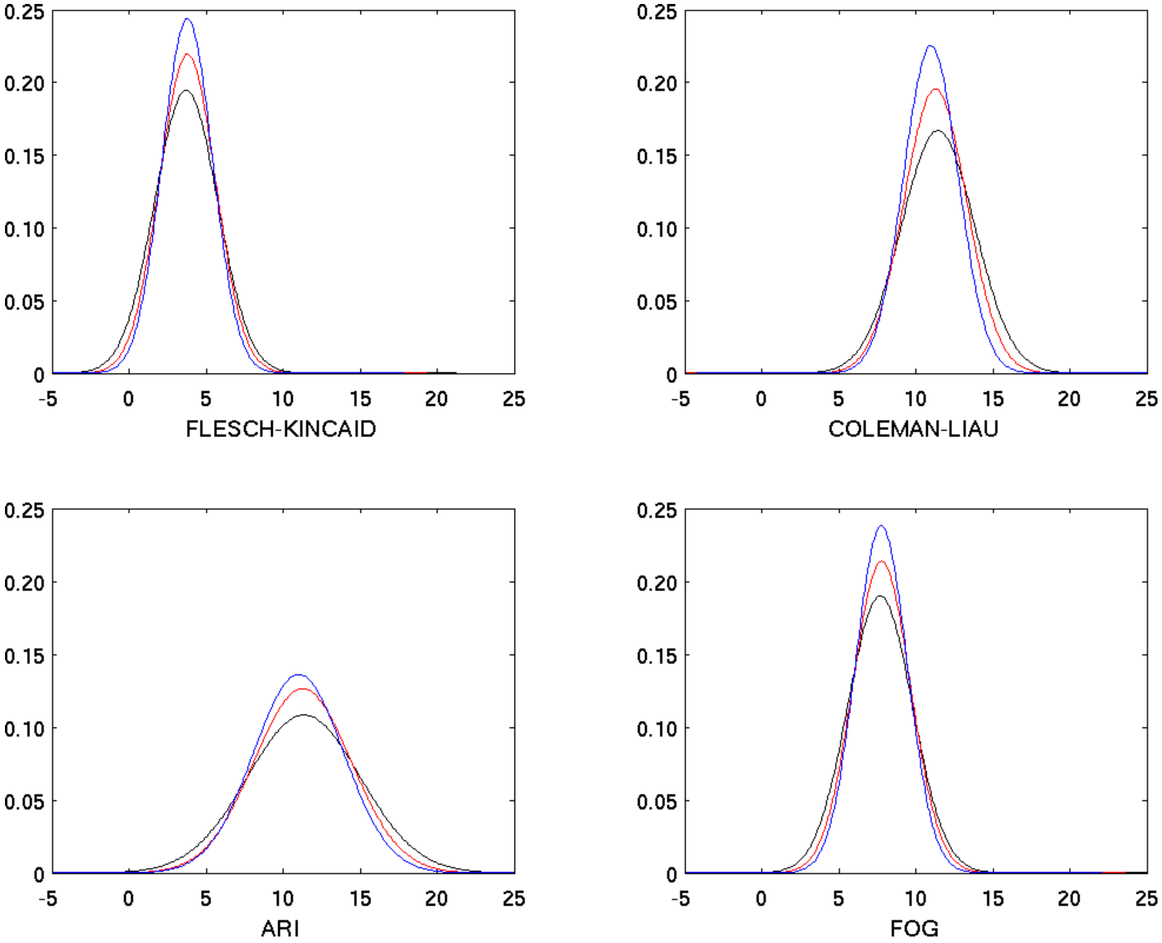

Figure 2: Gaussian Models of Flesch-Kincaid, Coleman-Liau, ARI, and FOG Measures Versus ILR Level for English Translations of Chinese Source Data (the results for levels 2, 2+, and 3 are in black, red, and blue, respectively)

6.0 EXPERIMENTAL RESULTS WITH LANGUAGE MODELING

This section presents the results of two types of experiments conducted on automatically detecting or identifying the ILR level of texts. Detection and identification experiments are similar in many respects, but they have subtle differences. Further, system performance is generally assessed using different techniques and metrics.

The first type of experiment dealt with the task of detecting whether a text was from a given ILR level or not. In other words, if one is interested in level 2+ materials, the system should determine whether a given text is level 2+ or not level 2+. For the experiments conducted here, the model for "not level 2+" was trained using files from levels 2 and 3. In detection experiments, there are two types of errors—namely, misses and false alarms. A miss occurs when a text is really from a given level, but the system says that it isn't; a false alarm occurs when a text is not from a given level, but the system says it is. These two types of errors are controlled by setting a threshold. Raising the detection threshold results in fewer false alarms but more misses, while lowering the threshold results in fewer misses but more false alarms. System performance in detection experiments is often presented in terms of a receiver operating characteristic (ROC) curve or a detection error trade-off (DET) curve (Martin et al., 1997). We use the DET curve in this report. A second type of performance measure is called the equal error rate (EER). The EER is the probability of a miss (or a false alarm) that results when one sets the detection threshold to obtain equal miss and false alarm probabilities.

The second type of experiment dealt with the task of identifying to which of three ILR levels (2, 2+, or 3) a given text belonged. In this type of experiment (called closed-set identification), a text was compared against all three level models, and the level of the best scoring model was assigned as the hypothesized level (no matter how close the other models scored). Performance in these closed-set identification experiments was assessed with confusion matrices.

The classifier used to detect (or identify) the ILR level of an article was a statistical language model (LM) built using the CMU/Cambridge toolkit[6] (Clarkson and Rosenfeld, 1997). For each ILR level, all training articles of that level were grouped together and either a unigram, bigram, or trigram LM was built using Witten-Bell discounting (Manning and Schütze, 1999; Witten and Bell, 1991) and zero cut-offs. Each test article was then evaluated against each of the ILR level models and a Background (BKG) model comprised of all the training data. A final score for a test article, A_i, against an ILR level model, M_j, was derived as:

$$\mathrm{score}\left(A_i, M_j\right) = \frac{\sum_{n=1}^{N_{i,j}} \log\left(\Pr\left(\text{n-gram}_n \mid M_j\right)\right) - \log\left(\Pr\left(\text{n-gram}_n \mid M_{BKG}\right)\right)}{N_{i,j}}$$

where $N_{i,j}$ is the number of n-grams in article A_i that exist in both model M_j and the background model, M_{BKG}. The final log-likelihood scores for the test articles against the various models where then used to generate DET curves for each ILR level. For the closed-set identification experiments, the highest log-likelihood score was chosen as the hypothesized level, and the hypothesized and hand-labeled levels were used to compute confusion matrices.

6.1 ILR Level Detection Results

Figures 3 and 4 show the DET curves for the three ILR levels (with the best performing language model for each) for the Arabic source data and the English translation of the Arabic,

[6] `http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html`

respectively. In both cases, the level 3 detectors had the best performance followed by level 2 detectors; the level 2+ detectors performed the worst in both cases. Generally, unigram or bigram language models performed the best. The first two rows of Table 9 show the EER's for these experiments. The best EER performance from these experiments was 16.0% for the level 3 detector built with bigram language models on the English translations of the Arabic source data.

Figures 5, 6, and 7 show the DET curves for the three ILR levels (with the best performing language model for each) for the original (*i.e.,* character-segmented) Chinese source data, the word-segmented Chinese source data, and the English translation of the Chinese, respectively. In all cases, the level 3 detectors had the best performance followed by level 2 detectors; the level 2+ detectors performed the worst in all cases. Again, unigram or bigram language models performed the best. The last three rows of Table 9 show the EER's for these experiments. The best EER performance from these experiments was 21.8% for the level 3 detector built with unigram or bigram language models on the English translations of the Chinese source data. There is less than a 1% EER difference in performance between the best detectors for the character-segmented and word-segmented data, with the character segmented data being the better of the two. While word segmentation provided no benefit in this case, a different Chinese word segmenter might yield improved performance. Also, word segmentation might provide some benefit on a different data set.

6.2 Closed-Set ILR Level Identification Results

Tables 10–19 show the confusion matrices that result from performing closed-set identification (CSID) of ILR level. Each table shows the reference (*i.e.,* true) ILR level in the first column and the system hypothesized ILR level in the second column. The grayed rows represent correct identification of the level. The overall results for any experiment are shown in the right-hand column (marked “Total”), while the individual columns under the topic headings show the breakout of the results for the tested files of the corresponding topics. For example, Table 10 shows the confusion matrix that results when using unigram models built on the Arabic source data. From Table 2, one can see that there are six Arabic documents that are labeled as level 2 and classified as “science” (denoted “sci”). Due to the round-robin training and testing procedure, each file is tested three times (against model sets built with different training file sets), so these six files result in 18 tests. One of these 18 tests results in a correct classification of being in level 2 (*i.e.,* the 1 in the “Ref 2 Hyp 2” row and “sci” column), while five of these tests result in misclassifications of being in level 2+ (*i.e.,* the “Ref 2 Hyp 2+” row and “sci” column) and the remaining twelve tests result in misclassifications of being in level 3.

Comparing the CSID results from using unigram models versus those from using bigram models, one can see that regardless of language (or use of the English translation or word segmentation in Chinese), overall level 3 results are dramatically better for unigram models than for bigram models. However, this trend is not consistent with the results for the EERs from the detection experiments shown in Table 9, where one can see that the EER differences are often small between unigram and bigram models and that sometimes the bigram models performed better than the unigram models. On the other hand, overall level 2 results in the CSID experiments are dramatically better for bigram models than for unigram models. Again, this trend contrasts with the results of the detection experiments, where the unigram models consistently outperformed the bigram models. It would be interesting to see if these same biases still hold with a larger database that was more evenly distributed across ILR level.

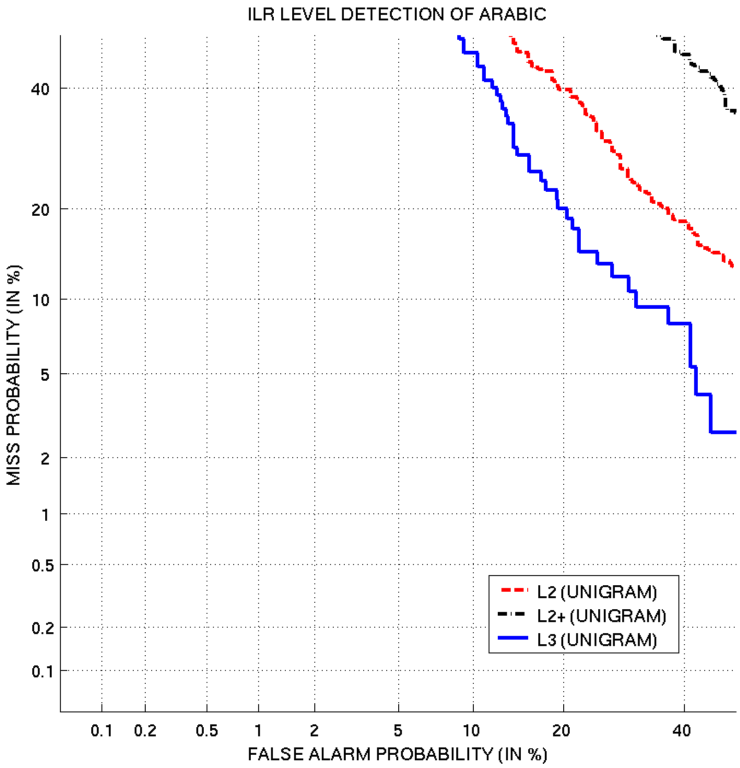

Figure 3: Detection Performance on Arabic Source Data

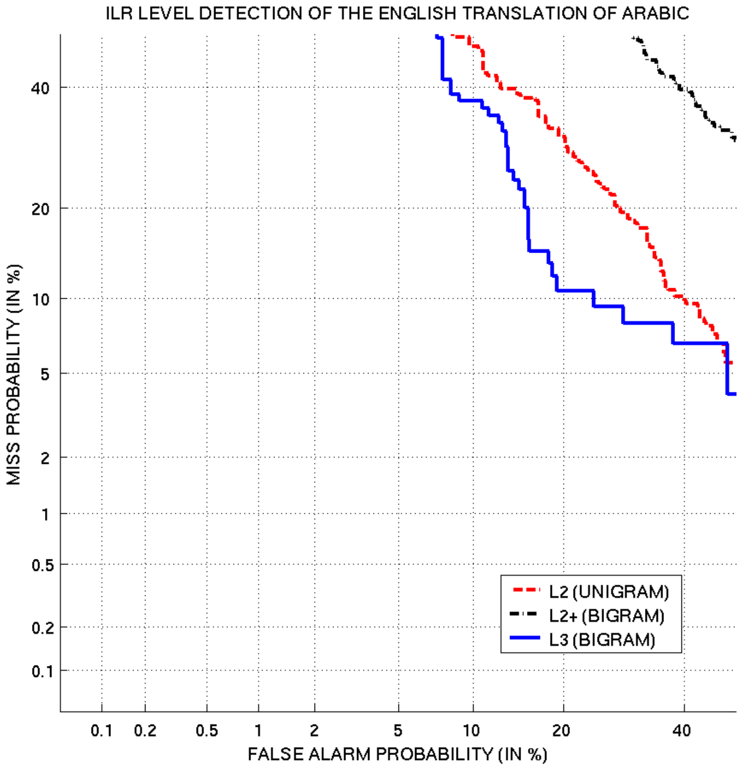

Figure 4: Detection Performance on English Translations of Arabic Source Data

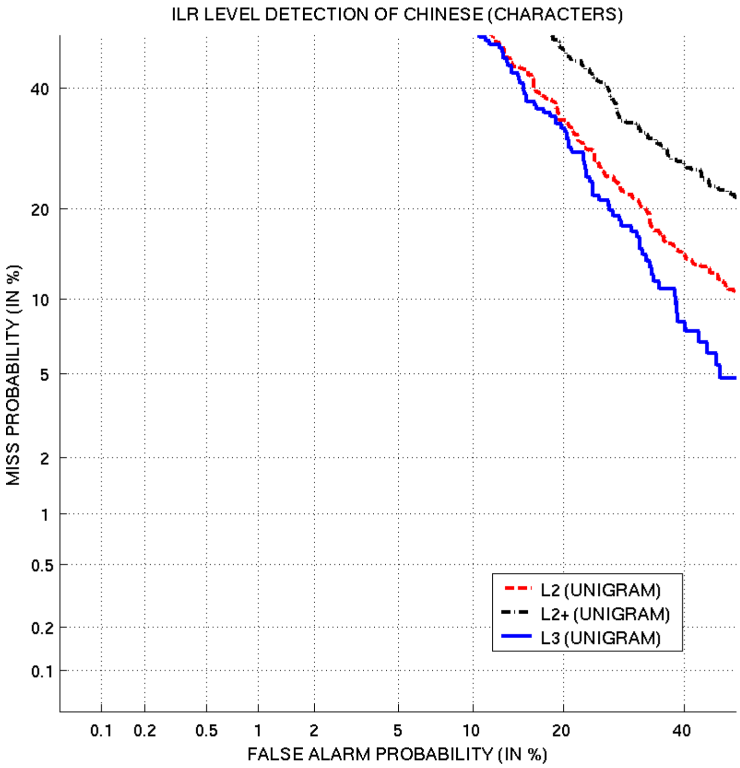

Figure 5: Detection Performance on Character-Segmented Chinese Source Data

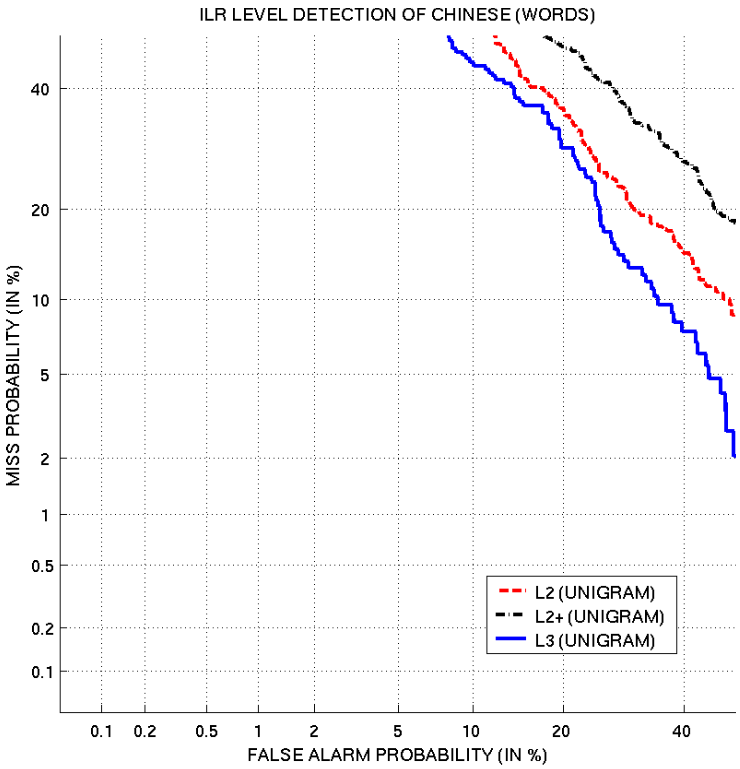

Figure 6: Detection Performance on Word-Segmented Chinese Source Data

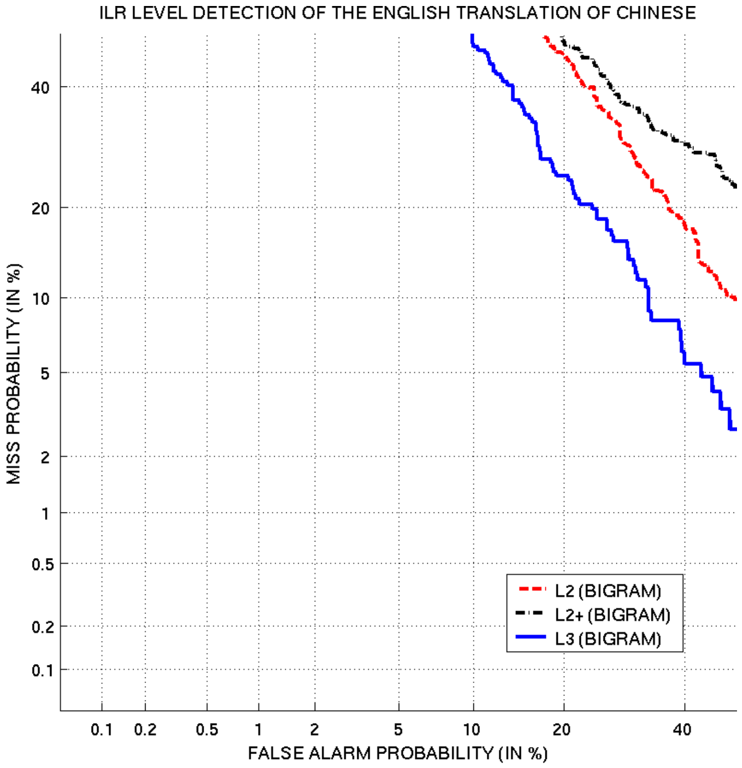

Figure 7: Detection Performance on English Translations of Chinese Source Data

Table 9: Equal Error Rates for the ILR Level Detectors on the Original and Translated Texts Using Unigram (U), Bigram (B), and Trigram (T) Language Models (boldfaced entries indicate the best language model type (U, B, or T) for each material type and ILR level)

Material	Level 2			Level 2+			Level 3		
	U	B	T	U	B	T	U	B	T
Arabic Source	**28.4**	33.3	35.9	**43.2**	49.4	51.9	**20.0**	33.3	34.7
English Translation of Arabic Source	**24.6**	26.1	27.8	40.7	**39.5**	39.5	17.3	**16.0**	21.3
Chinese Source (characters)	**25.6**	33.3	35.1	**32.1**	32.7	33.3	23.8	**22.4**	22.4
Chinese Source (word segmented)	**25.6**	32.0	32.5	**33.0**	33.7	36.8	23.8	**23.1**	29.3
English Translation of Chinese Source	**29.5**	29.6	31.4	36.2	**33.7**	37.8	**21.8**	21.8	26.5

One question that naturally arises given the uneven topic distribution across the ILR levels is whether the level detection results are really just a function of topic spotting. In breaking out the CSID results by topic, it was hoped that this question might be addressed at least to some degree; however, the topic break outs do not yield a clear answer to this question. For example, the results on the Arabic source data with unigram models shown in Table 10 indicate that level 2 articles in the society topic are very often misclassified as being from level 2+ or 3. Such a result could be consistent with topic spotting, given the large percentage of level 2+ and 3 articles also from the society topic. However, level 2 articles in the security topic are also very often misclassified as being from level 2+ or 3, yet levels 2+ and 3 have very few articles from the security topic. In general, if a given treatment (*i.e.,* language model order, Chinese word segmentation versus character segmentation, or English translation versus original source) improved (or worsened) the CSID performance for a given ILR level, then it tended to have that effect across topic; there do not appear to be any glaring differences between topics in this regard (at least in this experiment).

6.3 Detection Experiments with Random Shuffles

As previously mentioned, the distribution of topics across ILR level is not uniform for the database used in these experiments, so it is unclear to what extent the detection and classification performance results discussed in the previous two subsections are the result of ILR level detection/classification versus just topic detection/classification. Those experiments used the original data splits discussed in Subsection 4.2, which were fixed in an effort to balance topics as much as possible across the splits. However, due to the small size of the database, there were many topics that were not well balanced. While one could make a considerable effort to build a database consisting of a balance of topics and ILR levels, it is far more likely that any real database used for training would have an unbalanced set of topics and ILR levels. To get a better sense of the variability that one might encounter with a set of articles of unknown topic distribution, we used the Fisher-Yates shuffle to generate 100 training/testing lists as discussed in Subsection 4.2.

Figures 8–12 show the detection performance comparisons between the original data splits (solid lines) and those resulting from the Fisher-Yates shuffle (dashed lines) for the original Arabic source data, the English translations of the Arabic, the character-segmented Chinese source data, the word-segmented Chinese source data, and the English translations of the Chinese, respectively. Tables 20–24 show the corresponding comparisons in EERs between the original data splits and those from the Fisher-Yates shuffle for unigram, bigram, and trigram models for the various Arabic and Chinese conditions. In the tables, the column labeled "Combined" under the heading "Fisher-Yates Shuffle" indicates the EER obtained by concatenating the 100 sets of score files into a single combined score file, while the columns labeled "Mean" and "Std. Dev." under the heading "Fisher-Yates Shuffle" indicate the mean and standard deviation, respectively, of the EER obtained by considering the 100 sets of score files separately.

In general, the trends found with the Fisher-Yates shuffle were similar to those found with the original splits. Level 3 exhibited the best detection performance, while Level 2+ exhibited the worst performance. On the Arabic source data, the best performance on level 3 was obtained using unigram models, which yielded an EER of 16.6% when the score files were combined and a mean EER of 13.8% when the score files were considered separately. In the latter case, the standard deviation of the EERs was 10.5%, which was rather high relative to the standard deviations for some of the other ILR levels, but not unexpected due to the smaller number of level 3 articles available for building models. On the character-segmented Chinese source data,

Table 10: Confusion Matrix for Unigram Models on Arabic Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**4** **1.2%**	**9** **2.6%**	**6** **1.7%**	**12** **3.5%**	**11** **3.2%**	**5** **1.4%**	**1** **0.3%**	**14** **4.1%**	**5** **1.4%**	**2** **0.6%**	**69** **20.0%**
2	2+	6 1.7%	14 4.1%	4 1.2%	11 3.2%	4 1.2%	28 8.1%	5 1.4%	18 5.2%	22 6.4%	12 3.5%	124 35.9%
2	3	17 4.9%	22 6.4%	8 2.3%	13 3.8%	6 1.7%	21 6.1%	12 3.5%	13 3.8%	36 10.4%	4 1.2%	152 44.1%
Total												345
2+	2	0 0.0%	3 1.9%	0 0.0%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	4 2.5%
2+	**2+**	**0** **0.0%**	**16** **9.9%**	**1** **0.6%**	**0** **0.0%**	**0** **0.0%**	**16** **9.9%**	**0** **0.0%**	**3** **1.9%**	**22** **13.6%**	**2** **1.2%**	**60** **37.0%**
2+	3	12 7.4%	8 4.9%	5 3.1%	0 0.0%	0 0.0%	16 9.9%	6 3.7%	3 1.9%	47 29.0%	1 0.6%	98 60.5%
Total												162
3	2	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
3	2+	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 4.0%	0 0.0%	0 0.0%	2 2.7%	0 0.0%	5 6.7%
3	**3**	**9** **12.0%**	**3** **4.0%**	**0** **0.0%**	**0** **0.0%**	**3** **4.0%**	**18** **24.0%**	**0** **0.0%**	**3** **4.0%**	**31** **41.3%**	**3** **4.0%**	**70** **93.3%**
Total												75

Table 11: Confusion Matrix for Bigram Models on Arabic Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**10** **2.9%**	**28** **8.1%**	**11** **3.2%**	**22** **6.4%**	**16** **4.6%**	**21** **6.1%**	**7** **2.0%**	**32** **9.3%**	**34** **9.9%**	**14** **4.1%**	**195** **56.5%**
2	2+	5 1.4%	11 3.2%	1 0.3%	8 2.3%	2 0.6%	18 5.2%	3 0.9%	7 2.0%	17 4.9%	3 0.9%	75 21.7%
2	3	12 3.5%	6 1.7%	6 1.7%	6 1.7%	3 0.9%	15 4.3%	8 2.3%	6 1.7%	12 3.5%	1 0.3%	75 21.7%
Total												345
2+	2	2 1.2%	17 10.5%	4 2.5%	0 0.0%	0 0.0%	9 5.6%	4 2.5%	3 1.9%	16 9.9%	1 0.6%	56 34.6%
2+	**2+**	**4** **2.5%**	**8** **4.9%**	**0** **0.0%**	**0** **0.0%**	**0** **0.0%**	**19** **11.7%**	**0** **0.0%**	**1** **0.6%**	**18** **11.1%**	**2** **1.2%**	**52** **32.1%**
2+	3	6 3.7%	2 1.2%	2 1.2%	0 0.0%	0 0.0%	5 3.1%	2 1.2%	2 1.2%	35 21.6%	0 0.0%	54 33.3%
Total												162
3	2	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 2.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 2.7%
3	2+	2 2.7%	2 2.7%	0 0.0%	0 0.0%	0 0.0%	8 10.7%	0 0.0%	1 1.3%	9 12.0%	3 4.0%	25 33.3%
3	**3**	**7** **9.3%**	**1** **1.3%**	**0** **0.0%**	**0** **0.0%**	**3** **4.0%**	**11** **14.7%**	**0** **0.0%**	**2** **2.7%**	**24** **32.0%**	**0** **0.0%**	**48** **64.0%**
Total												75

Table 12: Confusion Matrix for Unigram Models on English Translations of Arabic Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**12** **3.5%**	**30** **8.7%**	**13** **3.8%**	**28** **8.1%**	**20** **5.8%**	**17** **4.9%**	**8** **2.3%**	**37** **10.7%**	**18** **5.2%**	**5** **1.4%**	**188** **54.5%**
2	2+	1 0.3%	11 3.2%	4 1.2%	4 1.2%	1 0.3%	19 5.5%	4 1.2%	4 1.2%	21 6.1%	11 3.2%	80 23.2%
2	3	14 4.1%	4 1.2%	1 0.3%	4 1.2%	0 0.0%	18 5.2%	6 1.7%	4 1.2%	24 7.0%	2 0.6%	77 22.3%
Total												345
2+	2	0 0.0%	8 4.9%	5 3.1%	0 0.0%	0 0.0%	3 1.9%	2 1.2%	0 0.0%	4 2.5%	1 0.6%	23 14.2%
2+	**2+**	**0** **0.0%**	**12** **7.4%**	**0** **0.0%**	**0** **0.0%**	**0** **0.0%**	**13** **8.0%**	**0** **0.0%**	**3** **1.9%**	**30** **18.5%**	**0** **0.0%**	**58** **35.8%**
2+	3	12 7.4%	7 4.3%	1 0.6%	0 0.0%	0 0.0%	17 10.5%	4 2.5%	3 1.9%	35 21.6%	2 1.2%	81 50.0%
Total												162
3	2	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 1.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 1.3%
3	2+	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 2.7%	0 0.0%	0 0.0%	1 1.3%	1 1.3%	4 5.3%
3	**3**	**9** **12.0%**	**3** **4.0%**	**0** **0.0%**	**0** **0.0%**	**3** **4.0%**	**18** **24.0%**	**0** **0.0%**	**3** **4.0%**	**32** **42.7%**	**2** **2.7%**	**70** **93.3%**
Total												75

Table 13: Confusion Matrix for Bigram Models on English Translations of Arabic Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**22** **6.4%**	**40** **11.6%**	**18** **5.2%**	**34** **9.9%**	**21** **6.1%**	**39** **11.3%**	**15** **4.3%**	**44** **12.8%**	**47** **13.6%**	**12** **3.5%**	**292** **84.6%**
2	2+	2 0.6%	5 1.4%	0 0.0%	2 0.6%	0 0.0%	11 3.2%	3 0.9%	1 0.3%	13 3.8%	5 1.4%	42 12.2%
2	3	3 0.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	4 1.2%	0 0.0%	0 0.0%	3 0.9%	1 0.3%	11 3.2%
Total												345
2+	2	3 1.9%	19 11.7%	5 3.1%	0 0.0%	0 0.0%	18 11.1%	6 3.7%	6 3.7%	21 13.0%	3 1.9%	81 50.0%
2+	**2+**	**4** **2.5%**	**7** **4.3%**	**1** **0.6%**	**0** **0.0%**	**0** **0.0%**	**12** **7.4%**	**0** **0.0%**	**0** **0.0%**	**34** **21.0%**	**0** **0.0%**	**58** **35.8%**
2+	3	5 3.1%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	3 1.9%	0 0.0%	0 0.0%	14 8.6%	0 0.0%	23 14.2%
Total												162
3	2	1 1.3%	0 0.0%	0 0.0%	0 0.0%	1 1.3%	5 6.7%	0 0.0%	0 0.0%	2 2.7%	1 1.3%	10 13.3%
3	2+	1 1.3%	1 1.3%	0 0.0%	0 0.0%	1 1.3%	8 10.7%	0 0.0%	0 0.0%	12 16.0%	1 1.3%	24 32.0%
3	**3**	**7** **9.3%**	**2** **2.7%**	**0** **0.0%**	**0** **0.0%**	**1** **1.3%**	**8** **10.7%**	**0** **0.0%**	**3** **4.0%**	**19** **25.3%**	**1** **1.3%**	**41** **54.7%**
Total												75

Table 14: Confusion Matrix for Unigram Models on Character-Segmented Chinese Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**46** **7.4%**	**69** **11.1%**	**76** **12.2%**	**29** **4.7%**	**32** **5.2%**	**30** **4.8%**	**23** **3.7%**	**30** **4.8%**	**72** **11.6%**	**36** **5.8%**	**443** **71.3%**
2	2+	45 7.2%	7 1.1%	7 1.1%	7 1.1%	4 0.6%	5 0.8%	9 1.4%	3 0.5%	37 6.0%	3 0.5%	127 20.5%
2	3	20 3.2%	2 0.3%	4 0.6%	0 0.0%	0 0.0%	7 1.1%	4 0.6%	0 0.0%	14 2.3%	0 0.0%	51 8.2%
Total												621
2+	2	9 2.9%	13 4.1%	9 2.9%	5 1.6%	9 2.9%	12 3.8%	6 1.9%	3 1.0%	17 5.4%	0 0.0%	83 26.3%
2+	**2+**	**36** **11.4%**	**23** **7.3%**	**13** **4.1%**	**6** **1.9%**	**18** **5.7%**	**9** **2.9%**	**19** **6.0%**	**3** **1.0%**	**27** **8.6%**	**10** **3.2%**	**164** **52.1%**
2+	3	21 6.7%	3 1.0%	2 0.6%	1 0.3%	0 0.0%	6 1.9%	11 3.5%	0 0.0%	19 6.0%	5 1.6%	68 21.6%
Total												315
3	2	0 0.0%	2 1.4%	0 0.0%	3 2.0%	2 1.4%	0 0.0%	2 1.4%	4 2.7%	12 8.2%	0 0.0%	25 17.0%
3	2+	7 4.8%	2 1.4%	0 0.0%	0 0.0%	0 0.0%	3 2.0%	6 4.1%	4 2.7%	16 10.9%	2 1.4%	40 27.2%
3	**3**	**11** **7.5%**	**5** **3.4%**	**0** **0.0%**	**0** **0.0%**	**1** **0.7%**	**24** **16.3%**	**7** **4.8%**	**4** **2.7%**	**26** **17.7%**	**4** **2.7%**	**82** **55.8%**
Total												147

Table 15: Confusion Matrix for Bigram Models on Character-Segmented Chinese Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**91** **14.7%**	**73** **11.8%**	**83** **13.4%**	**34** **5.5%**	**33** **5.3%**	**36** **5.8%**	**30** **4.8%**	**32** **5.2%**	**111** **17.9%**	**39** **6.3%**	**562** **90.5%**
2	2+	19 3.1%	5 0.8%	4 0.6%	2 0.3%	3 0.5%	4 0.6%	6 1.0%	1 0.2%	11 1.8%	0 0.0%	55 8.9%
2	3	1 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.3%	0 0.0%	0 0.0%	1 0.2%	0 0.0%	4 0.6%
Total												621
2+	2	38 12.1%	23 7.3%	13 4.1%	7 2.2%	8 2.5%	17 5.4%	21 6.7%	6 1.9%	29 9.2%	5 1.6%	167 53.0%
2+	**2+**	**24** **7.6%**	**16** **5.1%**	**11** **3.5%**	**4** **1.3%**	**19** **6.0%**	**7** **2.2%**	**14** **4.4%**	**0** **0.0%**	**31** **9.8%**	**10** **3.2%**	**136** **43.2%**
2+	3	4 1.3%	0 0.0%	0 0.0%	1 0.3%	0 0.0%	3 1.0%	1 0.3%	0 0.0%	3 1.0%	0 0.0%	12 3.8%
Total												315
3	2	6 4.1%	4 2.7%	0 0.0%	3 2.0%	3 2.0%	4 2.7%	10 6.8%	9 6.1%	29 19.7%	3 2.0%	71 48.3%
3	2+	6 4.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	7 4.8%	5 3.4%	3 2.0%	15 10.2%	3 2.0%	39 26.5%
3	**3**	**6** **4.1%**	**5** **3.4%**	**0** **0.0%**	**0** **0.0%**	**0** **0.0%**	**16** **10.9%**	**0** **0.0%**	**0** **0.0%**	**10** **6.8%**	**0** **0.0%**	**37** **25.2%**
Total												147

Table 16: Confusion Matrix for Unigram Models on Word-Segmented Chinese Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**41** **6.6%**	**61** **9.8%**	**70** **11.3%**	**26** **4.2%**	**22** **3.5%**	**20** **3.2%**	**16** **2.6%**	**30** **4.8%**	**69** **11.1%**	**26** **4.2%**	**381** **61.4%**
2	2+	31 5.0%	11 1.8%	7 1.1%	6 1.0%	11 1.8%	8 1.3%	9 1.4%	3 0.5%	26 4.2%	9 1.4%	121 19.5%
2	3	39 6.3%	6 1.0%	10 1.6%	4 0.6%	3 0.5%	14 2.3%	11 1.8%	0 0.0%	28 4.5%	4 0.6%	119 19.2%
Total												621
2+	2	5 1.6%	8 2.5%	9 2.9%	0 0.0%	8 2.5%	10 3.2%	3 1.0%	5 1.6%	13 4.1%	2 0.6%	63 20.0%
2+	**2+**	**33** **10.5%**	**24** **7.6%**	**11** **3.5%**	**5** **1.6%**	**19** **6.0%**	**10** **3.2%**	**19** **6.0%**	**1** **0.3%**	**23** **7.3%**	**7** **2.2%**	**152** **48.3%**
2+	3	28 8.9%	7 2.2%	4 1.3%	7 2.2%	0 0.0%	7 2.2%	14 4.4%	0 0.0%	27 8.6%	6 1.9%	100 31.7%
Total												315
3	2	0 0.0%	0 0.0%	0 0.0%	1 0.7%	0 0.0%	0 0.0%	2 1.4%	2 1.4%	4 2.7%	0 0.0%	9 6.1%
3	2+	3 2.0%	2 1.4%	0 0.0%	0 0.0%	0 0.0%	3 2.0%	3 2.0%	1 0.7%	11 7.5%	2 1.4%	25 17.0%
3	**3**	**15** **10.2%**	**7** **4.8%**	**0** **0.0%**	**2** **1.4%**	**3** **2.0%**	**24** **16.3%**	**10** **6.8%**	**9** **6.1%**	**39** **26.5%**	**4** **2.7%**	**113** **76.9%**
Total												147

Table 17: Confusion Matrix for Bigram Models on Word-Segmented Chinese Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**90** **14.5%**	**75** **12.1%**	**83** **13.4%**	**32** **5.2%**	**31** **5.0%**	**32** **5.2%**	**29** **4.7%**	**33** **5.3%**	**102** **16.4%**	**37** **6.0%**	**544** **87.6%**
2	2+	20 3.2%	3 0.5%	3 0.5%	4 0.6%	5 0.8%	6 1.0%	7 1.1%	0 0.0%	16 2.6%	2 0.3%	66 10.6%
2	3	1 0.2%	0 0.0%	1 0.2%	0 0.0%	0 0.0%	4 0.6%	0 0.0%	0 0.0%	5 0.8%	0 0.0%	11 1.8%
Total												621
2+	2	28 8.9%	14 4.4%	9 2.9%	4 1.3%	11 3.5%	12 3.8%	11 3.5%	6 1.9%	24 7.6%	3 1.0%	122 38.7%
2+	**2+**	**29** **9.2%**	**21** **6.7%**	**15** **4.8%**	**5** **1.6%**	**16** **5.1%**	**9** **2.9%**	**20** **6.3%**	**0** **0.0%**	**31** **9.8%**	**10** **3.2%**	**156** **49.5%**
2+	3	9 2.9%	4 1.3%	0 0.0%	3 1.0%	0 0.0%	6 1.9%	5 1.6%	0 0.0%	8 2.5%	2 0.6%	37 11.7%
Total												315
3	2	0 0.0%	3 2.0%	0 0.0%	3 2.0%	2 1.4%	2 1.4%	4 2.7%	7 4.8%	21 14.3%	2 1.4%	44 29.9%
3	2+	9 6.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 4.1%	10 6.8%	4 2.7%	20 13.6%	3 2.0%	52 35.4%
3	**3**	**9** **6.1%**	**6** **4.1%**	**0** **0.0%**	**0** **0.0%**	**1** **0.7%**	**19** **12.9%**	**1** **0.7%**	**1** **0.7%**	**13** **8.8%**	**1** **0.7%**	**51** **34.7%**
Total												147

Table 18: Confusion Matrix for Unigram Models on English Translations of Chinese Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**39** **6.3%**	**45** **7.2%**	**65** **10.5%**	**23** **3.7%**	**24** **3.9%**	**14** **2.3%**	**16** **2.6%**	**24** **3.9%**	**27** **4.3%**	**17** **2.7%**	**294** **47.3%**
2	2+	35 5.6%	19 3.1%	7 1.1%	4 0.6%	7 1.1%	13 2.1%	9 1.4%	4 0.6%	36 5.8%	13 2.1%	147 23.7%
2	3	37 6.0%	14 2.3%	15 2.4%	9 1.4%	5 0.8%	15 2.4%	11 1.8%	5 0.8%	60 9.7%	9 1.4%	180 29.0%
Total												621
2+	2	4 1.3%	4 1.3%	6 1.9%	0 0.0%	5 1.6%	2 0.6%	2 0.6%	6 1.9%	6 1.9%	0 0.0%	35 11.1%
2+	**2+**	**27** **8.6%**	**20** **6.3%**	**11** **3.5%**	**3** **1.0%**	**18** **5.7%**	**4** **1.3%**	**20** **6.3%**	**0** **0.0%**	**24** **7.6%**	**8** **2.5%**	**135** **42.9%**
2+	3	35 11.1%	15 4.8%	7 2.2%	9 2.9%	4 1.3%	21 6.7%	14 4.4%	0 0.0%	33 10.5%	7 2.2%	145 46.0%
Total												315
3	2	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
3	2+	3 2.0%	1 0.7%	0 0.0%	0 0.0%	0 0.0%	3 2.0%	2 1.4%	1 0.7%	3 2.0%	0 0.0%	13 8.8%
3	**3**	**15** **10.2%**	**8** **5.4%**	**0** **0.0%**	**3** **2.0%**	**3** **2.0%**	**24** **16.3%**	**13** **8.8%**	**11** **7.5%**	**51** **34.7%**	**6** **4.1%**	**134** **91.2%**
Total												147

Table 19: Confusion Matrix for Bigram Models on English Translations of Chinese Data

Ref	Hyp	cul	ecn	env	geo	mil	pol	sci	sec	soc	tec	Total
2	**2**	**86** **13.8%**	**71** **11.4%**	**86** **13.8%**	**34** **5.5%**	**35** **5.6%**	**35** **5.6%**	**34** **5.5%**	**33** **5.3%**	**99** **15.9%**	**36** **5.8%**	**549** **88.4%**
2	2+	17 2.7%	5 0.8%	1 0.2%	2 0.3%	1 0.2%	5 0.8%	2 0.3%	0 0.0%	18 2.9%	3 0.5%	54 8.7%
2	3	8 1.3%	2 0.3%	0 0.0%	0 0.0%	0 0.0%	2 0.3%	0 0.0%	0 0.0%	6 1.0%	0 0.0%	18 2.9%
Total												621
2+	2	20 6.3%	22 7.0%	12 3.8%	5 1.6%	14 4.4%	15 4.8%	19 6.0%	6 1.9%	29 9.2%	3 1.0%	145 46.0%
2+	**2+**	**32** **10.2%**	**14** **4.4%**	**11** **3.5%**	**6** **1.9%**	**13** **4.1%**	**8** **2.5%**	**17** **5.4%**	**0** **0.0%**	**27** **8.6%**	**10** **3.2%**	**138** **43.8%**
2+	3	14 4.4%	3 1.0%	1 0.3%	1 0.3%	0 0.0%	4 1.3%	0 0.0%	0 0.0%	7 2.2%	2 0.6%	32 10.2%
Total												315
3	2	1 0.7%	3 2.0%	0 0.0%	3 2.0%	0 0.0%	1 0.7%	6 4.1%	5 3.4%	21 14.3%	1 0.7%	41 27.9%
3	2+	9 6.1%	1 0.7%	0 0.0%	0 0.0%	1 0.7%	8 5.4%	9 6.1%	7 4.8%	17 11.6%	3 2.0%	55 37.4%
3	**3**	**8** **5.4%**	**5** **3.4%**	**0** **0.0%**	**0** **0.0%**	**2** **1.4%**	**18** **12.2%**	**0** **0.0%**	**0** **0.0%**	**16** **10.9%**	**2** **1.4%**	**51** **34.7%**
Total												147

the best performance on level 3 was obtained using trigram models, which yielded an EER of 22.7% when the score files were combined and a mean EER of 18.7% when the score files were considered separately. In the latter case, the standard deviation of the EERs was 6.4%.

Overall, the standard deviations for the various levels and language model orders tended to be higher for the Arabic data (or its translation) than for the Chinese data (or its translation). The standard deviations of the Arabic ranged from 6.2–10.5%, while those for the Chinese ranged from 4.5–7.4%. It is tempting to attribute these differences to the larger amount of training data available for the Chinese; however, there are confounding factors such as the differences between the two languages and the topic distributions that might also play a role.

In the original split experiments, the particular Chinese word segmenter that was used provided no benefit over simply using character segmentation. However, the results under the Fisher-Yates shuffle are slightly different as the Chinese word segmentation provided a small improvement over the character segmentation for the level 3 materials (of 0.8% and 2.0% for the combined and mean of the separate results, respectively). It might be worthwhile to investigate other Chinese word segmentation algorithms to see if they could provide additional benefit.

Finally, it is interesting to compare the results on the original source languages to the corresponding results on the English translations. For a given ILR level and split configuration (original splits, Fisher-Yates splits combined as one score file, or Fisher-Yates splits considered separately), the difference between the EER for the source data and the EER for the English translation of the source can be computed. For Arabic, the ranges of the differences are -0.4–3.8% for unigram models, 2.2–17.3% for bigram models, and 5.1–8.6% for trigram models. The poorer performance on the Arabic source data for the bigram and trigram models relative to that for the English translations is likely due to the morphological complexity of Arabic (Badawi et al., 2004; Mace, 1998) and the fact that these experiments have not used a morphological analyzer. Future experiments should be done using a morphological analyzer such as the MADA system described in (Roth et al., 2008; Habash and Rambow, 2005; Habash and Sadat, 2006) or the simple systems described in (Shen et al., 2007, 2008). The Count-Mediated Morphological Analysis (CoMMA) process described in (Shen et al., 2009) might be employed as well. In contrast to the Arabic, the ranges of the differences for the Chinese (character-segmented source) versus the English translations are similar for the various language model orders—namely, -4.1–3.5% for unigrams, -3.9–5.2% for bigrams, and -4.5–3.7% for trigrams.

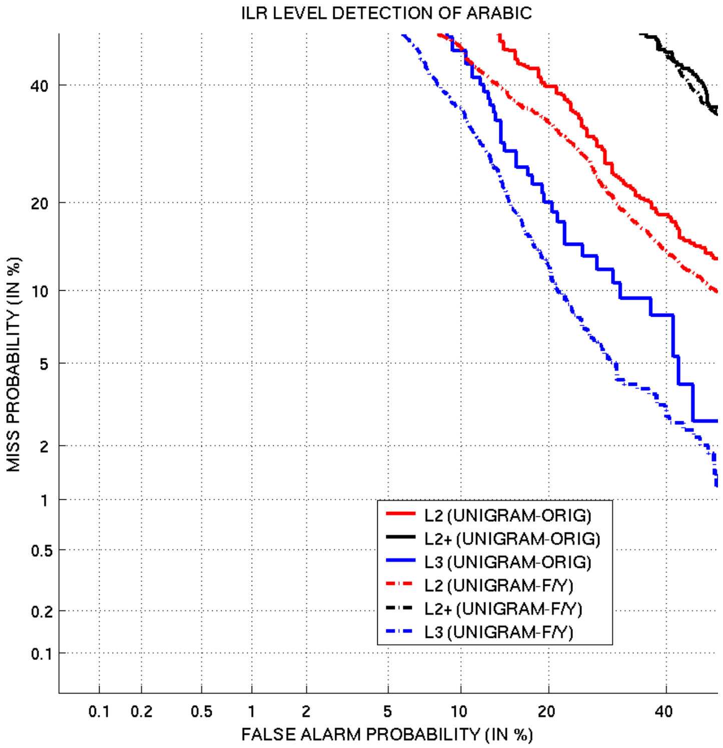

Figure 8: Comparison of Detection Performance on Arabic Source Data Using Original Splits (Solid Lines) and Fisher-Yates Shuffle (Dashed Lines)

Table 20: Comparison of Equal Error Rates on Arabic Source Data Using Original Splits and Fisher-Yates Shuffle with Unigram, Bigram, and Trigram Models

ILR Level	Original Splits	Fisher-Yates Shuffle		
		Combined	Mean	Std. Dev.
Unigram Models				
2	28.4%	26.3%	26.5%	6.2%
2+	43.2%	42.6%	40.9%	7.3%
3	20.0%	16.6%	13.8%	10.5%
Bigram Models				
2	33.3%	28.2%	29.0%	6.7%
2+	49.4%	44.9%	42.5%	8.8%
3	33.3%	23.2%	19.8%	6.6%
Trigram Models				
2	33.3%	33.2%	33.7%	6.7%
2+	46.9%	46.7%	45.7%	7.4%
3	29.3%	31.8%	25.2%	9.2%

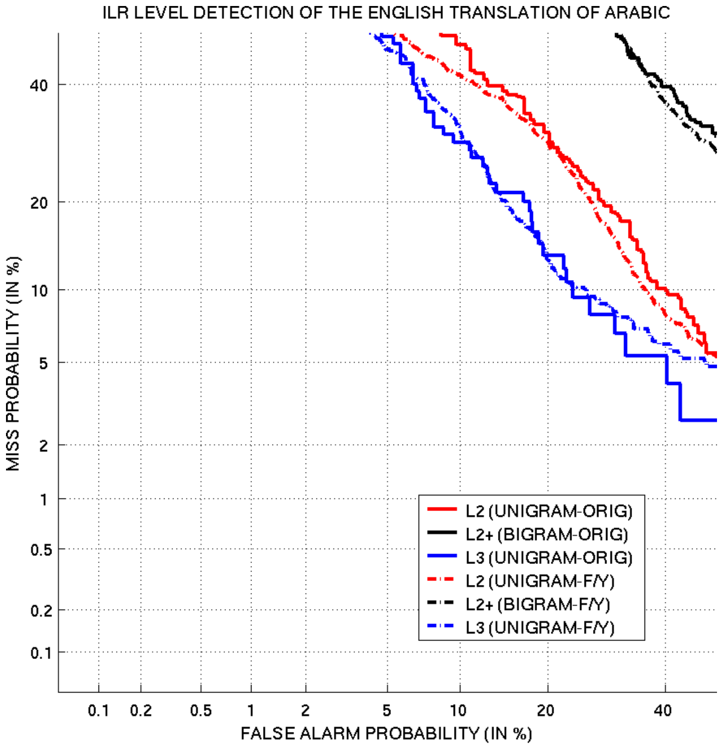

Figure 9: Comparison of Detection Performance on English Translations of Arabic Source Data Using Original Splits (Solid Lines) and Fisher-Yates Shuffle (Dashed Lines)

Table 21: Comparison of Equal Error Rates on English Translations of Arabic Source Data Using Original Splits and Fisher-Yates Shuffle with Unigram, Bigram, and Trigram Models

ILR Level	Original Splits	Fisher-Yates Shuffle		
		Combined	Mean	Std. Dev.
Unigram Models				
2	24.6%	23.9%	25.0%	6.3%
2+	40.7%	41.2%	38.4%	7.3%
3	17.3%	17.0%	12.4%	9.7%
Bigram Models				
2	26.1%	24.8%	26.8%	6.8%
2+	39.5%	38.5%	38.2%	8.0%
3	16.0%	18.8%	15.6%	8.3%
Trigram Models				
2	27.8%	26.6%	28.6%	6.9%
2+	39.5%	39.0%	39.4%	7.7%
3	21.3%	23.2%	19.2%	6.9%

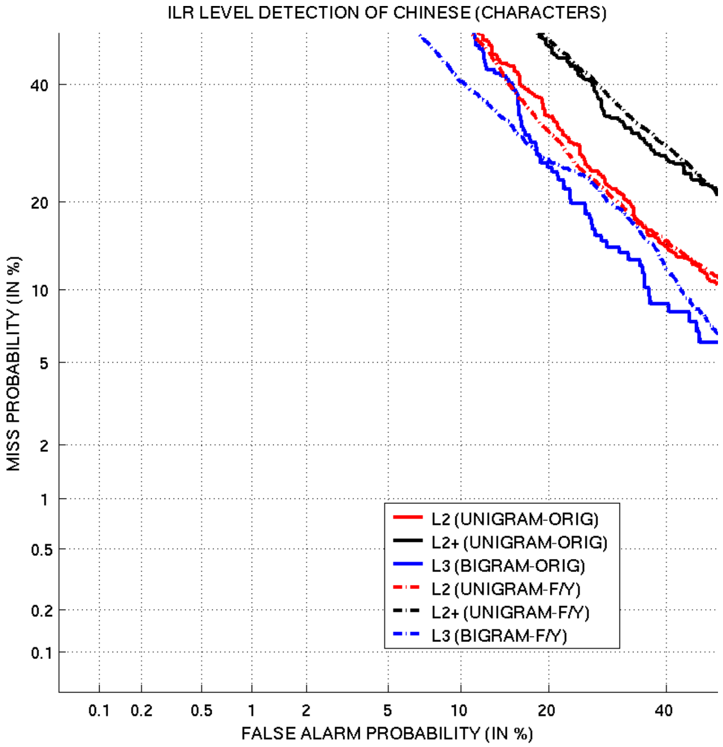

Figure 10: Comparison of Detection Performance on Character-Segmented Chinese Source Data Using Original Splits (Solid Lines) and Fisher-Yates Shuffle (Dashed Lines)

Table 22: Comparison of Equal Error Rates on Character-Segmented Chinese Source Data Using Original Splits and Fisher-Yates Shuffle with Unigram, Bigram, and Trigram Models

ILR Level	Original Splits	Fisher-Yates Shuffle		
		Combined	Mean	Std. Dev.
Unigram Models				
2	25.6%	24.8%	25.3%	4.7%
2+	32.1%	33.6%	31.6%	5.6%
3	23.8%	25.7%	22.6%	7.0%
Bigram Models				
2	33.3%	30.1%	30.9%	5.7%
2+	32.7%	33.3%	31.2%	5.6%
3	22.4%	24.4%	18.8%	7.0%
Trigram Models				
2	35.1%	34.8%	35.3%	5.5%
2+	33.3%	34.7%	33.3%	5.5%
3	22.4%	22.7%	18.7%	6.4%

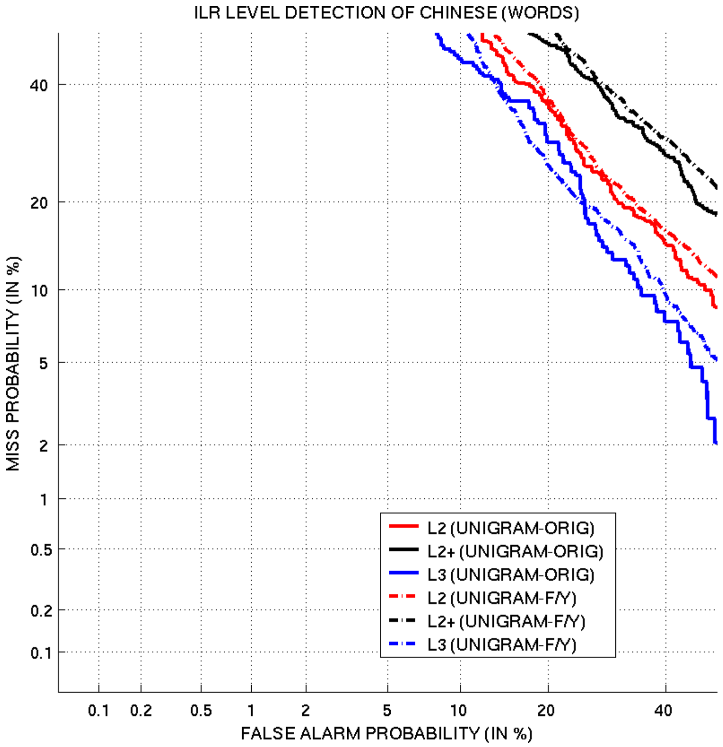

Figure 11: Comparison of Detection Performance on Word-Segmented Chinese Source Data Using Original Splits (Solid Lines) and Fisher-Yates Shuffle (Dashed Lines)

Table 23: Comparison of Equal Error Rates on Word-Segmented Chinese Source Data Using Original Splits and Fisher-Yates Shuffle with Unigram, Bigram, and Trigram Models

ILR Level	Original Splits	Fisher-Yates Shuffle		
		Combined	Mean	Std. Dev.
Unigram Models				
2	25.6%	26.6%	27.6%	4.7%
2+	33.0%	34.4%	33.6%	6.3%
3	23.8%	22.5%	19.4%	5.8%
Bigram Models				
2	32.0%	34.1%	34.7%	5.5%
2+	33.7%	34.8%	33.1%	5.8%
3	23.8%	20.7%	17.9%	6.2%
Trigram Models				
2	32.4%	35.1%	35.7%	5.4%
2+	37.8%	37.3%	36.0%	5.1%
3	25.9%	26.9%	24.4%	6.2%

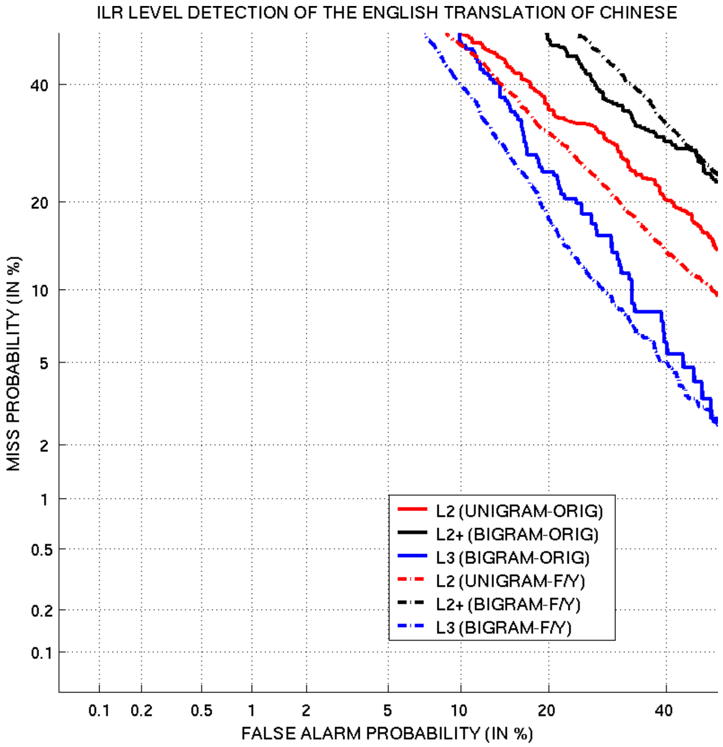

Figure 12: Comparison of Detection Performance on English Translations of Chinese Source Data Using Original Splits (Solid Lines) and Fisher-Yates Shuffle (Dashed Lines)

Table 24: Comparison of Equal Error Rates on English Translations of Chinese Source Data Using Original Splits and Fisher-Yates Shuffle with Unigram, Bigram, and Trigram Models

ILR Level	Original Splits	Fisher-Yates Shuffle		
		Combined	Mean	Std. Dev.
Unigram Models				
2	29.5%	25.3%	25.5%	4.5%
2+	36.2%	35.7%	34.1%	5.3%
3	21.8%	22.2%	19.2%	7.4%
Bigram Models				
2	29.6%	30.2%	30.8%	5.4%
2+	33.7%	36.8%	35.1%	5.4%
3	21.8%	19.2%	16.8%	5.8%
Trigram Models				
2	31.4%	34.3%	34.7%	5.1%
2+	37.8%	38.6%	36.6%	5.3%
3	26.5%	20.6%	18.9%	4.9%

7.0 CONCLUSIONS AND FUTURE WORK

This report has described experiments conducted on automatically determining the difficulty level of foreign language materials for the purpose of aiding teachers, students, and DoD linguists in finding suitable materials for supporting language learning and sustainment. The measure used as the indicator of difficulty was based on the ILR linguist proficiency scale. The experiments were conducted with a corpus of authentic Arabic and Mandarin Chinese materials from several genres that were hand-labeled for ILR level. The corpus contained materials at the 2, 2+, and 3 levels. ILR level detectors were built for these levels for both the original Arabic and Mandarin sources as well as for human-produced English translations of these sources. The detectors were based on statistical language modeling techniques. The EERs obtained ranged from 12.4–49.4% depending on the language, ILR level, language model order, and various other factors related to the experimental design. In general, the performance was best for discriminating level 3 materials from level 2 and 2+ materials, with EERs ranging from 12.4–33.3% across the languages (and translations), language model level, and experimental design. The performance was worst for discriminating level 2+ materials from level 2 and 3 materials, with EERs ranging from 31.2–49.4%.

There are a number of avenues for future research; however, the most important recommendation would be to collect and hand-label the ILR levels for a much larger collection of materials, especially in genres and languages of interest for a particular application. For example, if one wants to label broadcast new sources according to ILR level, then one should collect and label a sizable database of broadcast news materials for training and testing the detectors. The database considered in this work was of sufficient size to determine that the problem of detecting ILR level (at least for level 3 versus levels 2 and 2+) can potentially be addressed using statistical language modeling techniques, but any system meant for real use should be trained on an application-specific database. As seen in Subsection 6.3, the EERs for the various detectors often had rather large standard deviations, and these could be reduced with a larger database for training the detectors. Also, for any particular genre of data to be considered, it is important to determine the prior probabilities of the various ILR levels occurring as these prior probabilities, along with the costs of making miss or false alarm errors, are important for establishing the proper operating thresholds for the detectors.

In addition to collecting more training data, it would be interesting to investigate some of the grammar based features considered in (Schwarm and Ostendorf, 2005) and (Heilman et al., 2007). Parsers and part-of-speech taggers exist for both Arabic and Mandarin Chinese, so various grammar-based features could be examined. Also, as discussed in Subsection 6.3, the effects on the language model detector performance of applying an Arabic morphological analyzer to the Arabic source data should be investigated. Finally, as seen in the various experiments in Section 6.0 on the Chinese source data, the particular Chinese word segmenter that we used gave mixed results compared to building the language models on a character basis. However, there are other Chinese word segmenters available, and these should be investigated to determine if they can provide any benefit over the character-based language modeling.

BIBLIOGRAPHY

Badawi, E., Carter, M., and Gully, A., **Modern Written Arabic: A Comprehensive Grammar**, Routledge, London, 2004.

Chall, J., and Dale, E., **Readability Revisited: The New Dale-Chall Readability Formula**, Brookline Books/Lumen Editions, Brookline MA, May 1995.

Clarkson, P., and Rosenfeld, R., "Statistical language modeling using the CMU-Cambridge toolkit," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, September 1997, pp. 2707–2710.

Coleman, M., and Liau, T., "A computer readability formula designed for machine scoring," *Journal of Applied Psychology*, **60**(2), 1975, pp. 283–284.

Collins-Thompson, K., and Callan, J., "A language modeling approach to predicting reading difficulty," in *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, Boston MA, May 2004.

Cover, T., and Hart, P., "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, **IT-13**(1), January 1967, pp. 21–27.

Dale, E., and Chall, J., "A formula for predicting readability," *Educational Research Bulletin*, **27**, January and February 1948, pp. 1–20 and 37–54.

DuBay, W., **The Principles of Readability**, Impact Information, Costa Mesa CA, August 2004.

Durstenfeld, R., "Algorithm 235: Random permutation," *Communications of the ACM*, **7**, July 1964.

Fisher, R., and Yates, F., **Statistical Tables for Biological, Agricultural, and Medical Research**, Oliver & Boyd, London, third ed., 1948.

Flesch, R., "A new readability yardstick," *Journal of Applied Psychology*, **32**(3), June 1948, pp. 221–233.

Gunning, R., **The Technique of Clear Writing**, McGraw-Hill, New York, 1952.

Habash, N., and Rambow, O., "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Ann Arbor MI, June 2005.

Habash, N., and Sadat, F., "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New York NY, June 2006.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M., "Combining lexical and grammatical features to improve readability measures for first and second language texts," in *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, Rochester NY, April 2007.

Interagency Language Roundtable, "ILR Index," 3 March 2010, `http://www.govtilr.org`.

Joachims, T., "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the Tenth European Conference on Machine Learning*, Chemnitz, Germany, April 1998, pp. 137–142.

Joachims, T., "Making large-scale support vector machine learning practical," in Schölkopf, B., Burges, C., and Smola, A., (eds.) *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge MA, 1999, pp. 169–184.

Kincaid, J., Fishburne, Jr., R., Rogers, R., and Chissom, B., *Derivation of New Readability Formulas for Navy Enlisted Personnel*, Research Branch Report 8-75, Naval Technical Training, U.S. Naval Air Station, Memphis TN, 1975.

Kitson, H., **The Mind of the Buyer**, Macmillan, New York, 1921.

Knuth, D., **The Art of Computer Programming Vol. 2**, Addison-Wesley, Boston, third ed., 1998.

Lively, B., and Pressey, S., "A method for measuring the 'vocabulary burden' of textbooks," *Educational Administration and Supervision*, **9**, 1923, pp. 389–398.

Mace, J., **Arabic Grammar: A Reference Guide**, Edinburgh University Press, Edinburgh, 1998.

Manning, C., and Schütze, H., **Foundations of Statistical Natural Language Processing**, MIT Press, Cambridge MA, 1999.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, September 1997, pp. 1895–1898.

McLaughlin, G., "SMOG grading—A new readability formula," *Journal of Reading*, **12**(8), May 1969, pp. 639–646.

Mitchell, T., **Machine Learning**, McGraw-Hill, New York, 1997.

Rabin, A., "Determining difficulty levels of text written in languages other than english," in Zakaluk, B., and Samuels, S., (eds.) *Readability: Its Past, Present, and Future*, International Reading Association, Newark DE, 1988.

Roth, R., Rambow, O., Habash, N., Diab, M., and Rudin, C., "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking," in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Columbus OH, June 2008.

Schwarm, S., and Ostendorf, M., "Reading level assessment using support vector machines and statistical language models," in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Ann Arbor MI, June 2005, pp. 523–530.

Shen, W., Delaney, B., Aminzadeh, A., Anderson, T., and Slyh, R., "The MIT-LL/AFRL IWSLT-2009 MT system," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, December 2009, pp. 71–78.

Shen, W., Delaney, B., Anderson, T., and Slyh, R., "The MIT-LL/AFRL IWSLT-2007 MT system," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Trento, Italy, October 2007.

Shen, W., Delaney, B., Anderson, T., and Slyh, R., "The MIT-LL/AFRL IWSLT-2008 MT system," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Waikiki HI, October 2008, pp. 69–76.

Smith, E., and Senter, R., *Automated Readability Index*, Technical Report AMRL-TR-66-220, Aerospace Medical Research Laboratory, Air Force Systems Command, Wright-Patterson AFB OH, November 1967.

Spache, G., "A new readability formula for primary-grade reading materials," *Elementary School Journal*, **53**(7), 1953, pp. 410–413.

Vapnik, V., **The Nature of Statistical Learning Theory**, Springer, New York, 1995.

Vogel, M., and Washburne, C., "An objective method of determining the grade placement of children's reading material," *Elementary School Journal*, **28**, January 1928, pp. 373–381.

Witten, I., and Bell, T., "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory*, **37**(4), July 1991, pp. 1085–1094.

LIST OF ACRONYMS

ARI	Automated Readability Index (a readability measure)
BKG	background (model)
CLI	Coleman-Liau Index (a readability measure)
CMU	Carnegie Mellon University
CoMMA	Count-Mediated Morphological Analysis (system)
CSID	closed-set identification
DCRI	Dale-Chall Readability Index (a readability measure)
DET	detection error trade-off
DoD	Department of Defense
EER	equal error rate
FKGL	Flesch-Kincaid Grade Level (a readability measure)
FOG	Gunning Fog Grade (a readability measure)
FRE	Flesch Reading Ease (a readability measure)
GLOSS	Global Language Online Support System
GNU	GNU's Not Unix (a recursive acronym for a free Unix-like operating system)
Hyp	hypothesis
ILR	Interagency Language Roundtable (also denotes the ILR linguist proficiency scale)
kNN	k-Nearest Neighbor (a classification algorithm)
L2	Level 2 on the ILR proficiency scale
L2+	Level 2+ on the ILR proficiency scale
L3	Level 3 on the ILR proficiency scale
LDC	Linguistic Data Consortium
LM	language model
MADA	Morphological Analysis and Disambiguation for Arabic (system)
Ref	reference
ROC	receiver operating characteristic
SBAR	a clause introduced by a (possibly empty) subordinating conjunction
SG	Spache Grade (a readability measure)
SMOG	Standard Measure of Gobbledygook (a readability measure)
US	United States