Building Parallel Monolingual Gan Chinese Dialects Corpus

Fan XU, Mingwen WANG, Maoxi LI
School of Computer Information Engineering, Jiangxi Normal University
Nanchang 330022, China
{xufan, mwwang, mosesli}@jxnu.edu.cn

Abstract

Automatic language identification of an input sentence or a text written in similar languages, varieties or dialects is an important task in natural language processing. In this paper, we propose a scheme to represent Gan (Jiangxi province of China) Chinese dialects. In particular, it is a two-level and fine-grained representation using Chinese character, Chinese Pinyin and Chinese audio forms. Guided by the scheme, we manually annotate a Gan Chinese Dialects Corpus (GCDC) including 131.5 hours and 310 documents with 6 different genres, containing news, official document, story, prose, poet, letter and speech, from 19 different Gan regions. In addition, the preliminary evaluation on 2-way, 7-way and 20-way sentence-level Gan Chinese Dialects Identification (GCDI) justifies the appropriateness of the scheme to Gan Chinese dialects analysis and the usefulness of our manually annotated GCDC.

Keywords: Parallel Corpus, Monolingual, Gan Chinese Dialects

1. Introduction

Automatic language identification of an input sentence or a document is an important task in Natural Language Processing (NLP), especially when processing speech or social media messages. Currently, the interest in language resources and its computational models for the study of similar languages, varieties and dialects has been growing substantially in the last few years (Zampieri et al, 2014, 2015, 2017; Malmasi et al., 2016). Meanwhile, an increasing number of dialect corpus and corresponding computational models have been released for Catalan, Russian, Slovene, etc. However, no free corpus has been released for the similar, varieties or dialects in Mandarin Chinese. Therefore, in this paper, we focus on the corpus building and its computational model design for the closely related parallel monolingual Gan Chinese languages.

As we all know, Chinese is spoken in different regions, with distinct differences among regions. There are different expressions for a same concept among the closely related Gan Chinese languages, varieties and dialects. For example, 今里 jin li 'today', 今家 jin jia 'today', 今宁 jin ning 'today', 今兜 jin dou 'today', 今朝 jin zhao 'today' are the valid expressions in Nanchang, Yichun, Jian, Fuzhou and Yingtan district in Jiangxi province (Gan in short) of China, respectively. Although these expressions are different, they have the same semantic meanings. They all refer to 今天 jin tian 'today' in Mandarin Chinese (called 普通话 Putonghua 'common language' in Mainland China).

More specifically, firstly, we present a scheme to handle Gan Chinese dialects which is a fine-grained representation using Chinese character, Chinese Pinyin and Chinese audio forms. Based on the scheme, we manually annotate a parallel Gan Chinese Dialects Corpus (GCDC) consists of 310 documents with 6 different genres (news, official document, story, prose, poet, letter and speech) from 19 different Gan regions. As a byproduct, the corpus contains the parallel Gan Chinese audio with 131.5 hours. Besides, we conduct a preliminary experiment on the proposed GCDC through the sentence-level Gan Chinese Dialects Identification (GCDI) task. The simple but effective character Chinese Pinyin uni-gram yields a strong baseline on 2-way, 7-way and 20-way Gan dialects discrimination. The overall accuracy can reach to 78.64% on the fine-grained 20-way classification, which shows the automatic Gan Chinese dialects identification should be feasible. The evaluation result justifies the appropriateness of the scheme to Gan Chinese dialects analysis and the usefulness of our manually annotated GCDC.

The rest of this paper is organized as follows. Section 2 overviews related work. In Section 3, we present the scheme to deal with Gan Chinese dialects. Section 4 describes the annotation and an annotation instance of the GCDC. In Section 5, we present our preliminary experiment for the sentence-level Gan Chinese dialects identification on the proposed GCDC, and we conclude this work in Section 6 and present future directions.

2. Related Work

In this section, we describe the representative dialect corpus and its corresponding discrimination models.

2.1 Parallel Corpus

In the past decade, several parallel corpora among different languages have been proposed, e.g. Chinese-English (Ayan and Dorr, 2006), Japanese-English (Takezawa et al., 2002) and French-English (Mihalcea and Pedersen, 2003). They are annotated either at word-level or phrase-level alignment between two different languages (bilingual). Recently, many researchers pay attention to the parallel corpora only in the closely related languages (monolingual), varieties and dialects (Zampieri et al, 2014, 2015, 2017; Malmasi et al., 2016) which containing Bulgarian, Macedonian, etc. and a group containing texts written in a set of other languages. However, none parallel corpora in the closely related languages in the Gan dialects has been freely released so far. The representative certain scale parallel corpora is the Greater China Region (GCR) corpora (Xu et al., 2015) which focus on Mandarin with simplified and traditional scripts.

2.2 Dialects Identification Models

Generally speaking, language identification among different languages is a task that can be solved at a high accuracy. For example, Simoes et al. (2014) achieved 97% accuracy for discriminating among 25 unrelated languages. However, it is generally difficult to distinguish between related languages or variations of a specific language (see Zampieri et al, 2014, 2015 for example). To

be more specific, Ranaivo-Malancon (2006) proposed features based on frequencies of character n-grams to identify Malay and Indonesian. Zampieri and Gebre (2012) found that word uni-grams gave very similar performance to character n-gram features in the framework of the probabilistic language model for the Brazilian and European Portuguese language discrimination. Tiedemann and Ljubesic (2012); Ljubesic and Kranjcic (2015) showed that the Na ̈ve Bayes classifier with uni-grams achieved high accuracy for the South Slavic languages identification. Grefenstette (1995); Lui and Cook (2013) found that bag-of-words features outperformed the syntax or character sequences-based features for the English varieties. Besides these works, other recent studies include: Spanish varieties identification (2014), Arabic varieties discrimination (Elfardy and Diab,2013; Zaidan and Callison-burch, 2014; Salloum et al.,2016; Tillmann et al.,2014) and Persian and Dari identification (Malmasi and Dras, 2015); Indian languages identification (Murthy and Kumar, 2006).

3. Annotation Scheme

In this section, we present the scheme to Gan Chinese dialects which has two-level partitions and three forms.

3.1 Two-level Partitions

Gan Chinese is spoken in different regions in Jiangxi province of China, with distinct differences between two regions. To be specific, Table 2 shows a two-level Gan dialects partition is provided. The first level contains six regions of Gan dialects (Yan Sen, 1986), such as 昌靖片 'chang jing region', 宜萍片 'yi ping region', 吉莲片 'ji lian region', 抚广片 'fu guang region', 鹰弋片 'yin yi region', 客家话 'Hakka'. The six regions are further divided into 19 sub-regions in the second level. For example, 昌靖片 'chang jing region' contains 5 sub-regions, such as 新建 'xinjian', 南昌 'nanchang' and so on.

3.2 Three Forms

Chinese Pinyin: Pinyin is basically the alphabet for the Chinese language. The Pinyin system was invented to help people pronounce the sound of the Chinese characters. It is a Romanization system used to learn Mandarin. It transcribes the sounds of Mandarin using the Western (Roman) alphabet. It is well know that pronunciation is vital for any language. Therefore, we annotate Chinese Pinyin into our corpus. Figure 1 and Table 1 show pitch contours of lexical tones in Mandarin (Chen et al., 2016). In our corpus annotation, we annotate the pitch height with 1, 2, 3 and 4 accordingly.

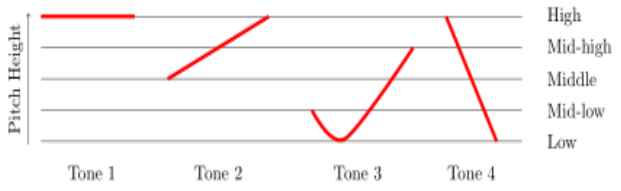

Figure 1: Pitch contours of lexical tones in Mandarin. The vertical axis denotes pitch height, whereas the horizontal staves indicate different tone levels within one's comfortable vocal range.

Tone	Pitch Contour	English Equivalent
1	High-level	Singing
2	High-rising	Question-final intonation; e.g., What?
3	Dipping	No equivalent; e.g., nǐhǎo, hello
4	Falling	Curt commands; e.g., Stop!

Table 1: Lexical Tones in Mandarin.

Chinese character: We observe that the same concept can be expressed using different linguistic expressions for the different region of Gan dialects as mentioned in the Introduction section.

Chinese audio: Furthermore, we present Chinese audio as a byproduct in this corpus. It consists of the parallel Gan Chinese audio and mandarin Chinese sound for each document.

4. Gan Chinese Dialects Corpus

In this section, we address the key issues with the GCDC annotation.

4.1 Annotator Training

The annotator team consists of a Ph.D. in Chinese linguistics as the supervisor (senior annotator) and 19 undergraduate students from different 19 Gan regions in Chinese linguistics as annotators. An annotator of a given region works only in data of his/her area. The annotation is done in four phases. In the first phase, the annotators spend 1 month on learning the principles of scheme. In the second phase, the annotators spend 1 month on independently annotating the same 30 documents, and another 1 month on crosschecking to resolve the difference and to revise the guidelines. In the third phase, the annotators spend 2 months on annotating the remaining 280 documents. In the final phase, the supervisor spends 1 month carefully proofread all 310 documents.

4.2 Corpus Statistics

Currently, the GCDC corpus consists of the representative 19 sub-regions of Gan dialects and their statistics as shown in Table 2. Given the above scheme, we annotate parallel 310 XML-style documents with 6 different genres (news, official document, story, prose, poet, letter and speech), containing 218 newswire documents from Chinese Treebank 6.0 with Linguistic Data Consortium (LDC) catalog number LDC2007T36, and other 92 documents for the remaining genres from the internet using Baidu search engine with official document, story, prose, poet, letter and speech as key words. Specifically, we don't have parallel sentences for each variant of each sentence in all documents, and the documents included differ among the dialects but are all parallel with respect to a Mandarin translation. We require the annotators to annotate the documents included differ among the dialects but are all parallel with respect to a Mandarin translation. As a byproduct, it has the 131.5-hour audios, wherein 69.0 hours Gan dialects sound and 62.50 hours Putonghua sound, and the total number of sentence in the corpus is 3,878. Table 2 shows the statistics in detail with the number of non-news genre are shown in parentheses.

Dialects region (level-1)	Dialects location (level-2)	Number of document	Number of sentence
昌靖片 chang jing region	新建 xinjian	10(4)	-
	南昌 nanchang	10(3)	-
	湖口 hukou	10(2)	-
	都昌 douchang	6(6)	-
	靖安 jingan	16(0)	-
	#total	52(15)	353(101)
吉莲片 ji lian region	吉安 jian	10(2)	-
	吉水 jishui	8(6)	-
	永丰 yongfeng	21(12)	-
	#total	39(20)	362(138)
抚广片 fu guang region	进贤 jingxian	10(2)	-
	东乡 dongxiang	14(2)	-
	抚州 fuzhou	17(12)	-
	#total	41(16)	230(105)
宜萍片 yi ping region	丰城 fengcheng	10(4)	-
	宜丰 yifeng	5(3)	-
	萍乡 pingxiang	7(5)	-
	#total	22(12)	165(99)
鹰弋片 yin yi region	余干 Yugan	10(4)	-
	上饶 shangrao	8(8)	-
	#total	18(12)	110(63)
客家话 Hakka	赣州 ganzhou	24(12)	-
	兴国 xinguo	12(2)	-
	大余 dayu	10(3)	-
	#total	46(17)	294(120)
普通话 Putonghua		156(92)	1113(625)
#total		310	3878

Table 2: Corpus statistics.

4.3 Quality Assurance

It is very challenging to check the agreement between annotators. We focus on 鹰弋片'yin yi region', and require another 2 annotators from this region to annotate 30 documents, 173 sentences, from the corpus. We calculate the annotation consistency value which is 0.93 for Pinyin. Due to the homophone phenomenon is obvious in Chinese character, we don't calculate the agreement for Chinese character. The high inter-annotator consistency in Chinese Pinyin guarantees the corpus's quality.

4.4 An Annotation Instance

Table 3 describes an annotation instance of GCDC for clarity.

```
<?xml version="1.0" encoding="GB2312" ?>
<document>
<voiceInfo>
  <Region>昌靖片 chang jing region</Region>
  <Location>新建 xinjian </Location>
  <Sex>女 Female</Sex>
  <Age>19</Age>
  <Genre>新闻 news </Genre>
  <Chanel>手机 mobile phone</Chanel>
  <FangyanTime>38 seconds </FangyanTime>
  <PutongTime>36 seconds</PutongTime>
  <FangyanFile>chtb_2946_fangyan.wav</FangyanFile>
  <PutongFile>chtb_2946_putong.wav</PutongFile>
</voiceInfo>
<sentence count="1">
  <putongContent>据 报道 ： 星期六 印度 和 巴基斯坦 军队 ， 在 科什米尔 停火线 一带 又 发生 了 新的 冲突 。According to a report, new conflicts in Kashmir ceasefire area were occurred between India and Pakistan on Saturday.
  </putongContent>
  <ganContent>居 报倒 ： 星期六 印度 和 巴基斯坦 军队 ， 赖 科什米尔 停我线 一带 又 发生 的 新个 冲突 。According to a report, new conflicts in Kashmir ceasefire area were occurred between India and Pakistan on Saturday.
  </ganContent>
  <putongPinyin>Ju4 Bao4Dao3 : Xing1Qi2Liu4 Yin4Du4 He2 Ba1Ji1Si1Tan3 Ju1Dui4 , Zai4 Ke1Shen2Mi3Er3 Ting2Huo3Xian4 Yi2Dai4 You4 Fa1Sheng1 Le1 Xin1De1 Chong1Tu1 .
  </putongPinyin>
  <ganPinyin>Ju4 Bao4Dao3 : Xing1Qi2Liu4 YIN4Du4 He2 Ba1Ji1Si1Tan3 Ju1Dui4 , Lai4 Ke1Shen2Mi3Er3 Ting2Wo3Xian4 Yi2Dai4 You4 Fa1Sheng1 De4 Xin1Ge4 Chong1Tu1 .
  </ganPinyin>
</sentence>
<sentence count="2">
  <putongContent>巴基斯坦 方面 说 ： 最近 发生 在 平泊尔 地区 的 冲突 中 ， 有 5 名 印度 士兵 被 打死 ， 很多 士兵 被 打伤 。It was reported by Pakistan that five Indian soldiers were killed and many soldiers were wounded in recent clashes in Pingboer area.
  </putongContent>
  <ganContent>巴基斯坦 方面 挖 ： 最将 发生 赖 平泊尔 那里 个 冲突 里面 ， 有 5 个 印度 当兵个 被 打死的 ， 好多 当兵个 被 打伤的 。It was reported by Pakistan that five Indian soldiers were killed and many soldiers were wounded in recent clashes in Pingboer area.
  </ganContent>
  <putongPinyin>Ba1Ji1Si1Tan3 Fang1Mian4 Shuo1 : Zui4Jin4 Fa1Sheng1 Zai4 Ping2Bo2Er3 Di4Qu1 De1 Chong1Tu1 Zhong1 , You3 Wu3 Ming2 Yin4Du4 Shi4Bing1 Bei4 Da3Si3 , Hen3Duo1 Shi4Bing1 Bei4 Da3Shang1.
  </putongPinyin>
  <ganPinyin>Ba1Ji1Si1Tan3 Fang1Mian4 Wa1 : Zui4Jiang1 Fa1Sheng1 Lai4 Ping2Bo2Er3 Na4Li3 Ge4 Chong1Tu1 Li3Mian4 , You3 En3 Ge4 Yin4Du4 Dang11Bing1Ge4 Bei4 Da3Si3De1 , Hao3Duo1 Dang11Bing1Ge4 Bei4 Da3Shang1De1 .
  </ganPinyin>
</sentence>
</document>
```

Table 3: An annotation instance for Gan dialects.

The example comes from file chtb2946 of CTB (Chinese Tree Bank) released by the LDC. The <voiceInfo> section describes the detail information, such as region, location, sex and age of annotator, genre type, record channel, duration and file. The <sentence> section demonstrates the specific contents including Chinese character and Chinese Pinyin, containing <putongContent> section refers to the Chinese character in Mandarin, <ganContent> section represents the Chinese character in specific Gan dialect. <putongPinyin> section indicates the Chinese Pinyin in Mandarin, while <ganPinyin> section means the Chinese Pinyin in specific Gan dialect. The whole corpus are available through the LREC 2018 repository.

5. Preliminary Experimentation

As mentioned in the Introduction section, automatic language identification of an input text is an important task in Natural Language Processing (NLP) because somebody must determine the language of the text before applying tools trained on specific language. For the sentence-level language identification, a user is given a single sentence, and the user needs to identify the language. Below, we recast the sentence-level dialects identification in the Gan dialects as a multi-class classification problem. Firstly, we will describe some features. Then, these features are fed into a classifier to determine the dialect of a sentence.

5.1 Features

In this section, we represent the character-level N-gram features.

Chinese Character Pinyin N-gram: According to the related work (Nikola and Denis, 2015; Cagri and Taraka, 2016), n-grams with $n \leq 3$ are effective features for discriminating general languages. Also, Cagri and Taraka (2016) showed their simple linear SVM model with n-gram feature is quite useful and hard to beat by current neural network models. Compared with English, no space exists between words in Chinese sentence. Therefore, we use character uni-grams, bi-grams and tri-grams in Chinese Pinyin as features. We take Pinyin with lexical tones or without it as different two kinds of features.

Chinese Character N-gram: While our corpus provides both Chinese character and Chinese Pinyin simultaneously, we also present Chinese Character uni-grams, bi-grams and tri-grams as features. This is because sometime Pinyin is not available in a specific situation.

5.2 Classifier and Evaluation Metric

Classifier: After extracting the above proposed features, we train a single multi-class linear kernel support vector machine using LIBLINEAR (Fan et al., 2008) for Gan Chinese dialects identification. They adopt the default parameters such as verbosity level with 1, trade-off between training error and margin with 0.01, slack rescaling, zero/one loss.

Evaluation Metric: We report system's performance using accuracy, which is the ratio of the number of the correctly predicted sentence divided by the total number of sentence for Gan dialects.

For the Gan dialects dataset, we generate three scenarios using 5-fold cross validation:

(1) **2-way detection**: We try to distinguish between two groups of dialects, the ones is 普通话 'Putonghua', and the others are the left 19 sub-regions of Gan dialects;

(2) **7-way detection**: The level-1 Gan dialects of 昌靖片 'chang jing region', 吉莲片 'ji lian region', 抚广片 'fu guang region', 宜萍片 'yi ping region', 鹰弋片 'yin yi region', 客家话 'Hakka' and 普通话 'Putonghua' as shown in Table 2 are considered;

(3) **20-way detection**: We detect both Mandarin and other level-2 19 Gan dialects of 新建 xinjian, 南昌 nanchang, 湖口 hukou, etc. as shown in Table 2 are all considered.

5.3 Experimentation Results

In this section, we report the experiment results for the Gan Chinese dialects identification on our dataset.

5.3.1 Results on Chinese Character Pinyin

Table 4 shows the performance on Chinese character Pinyin feature. As can be seen, the character uni-gram Pinyin feature yields the best performance on both news and other types of genres. Obviously, the performance of 2-way classification is higher than both 7-way and 20-way language discrimination. Strangely, the performance of the more fine-grained 20 different dialect labels task achieves higher results than the identification of only 7 labels. We attribute it to the parallel nature of the corpus. Basically, the performance is increased with the increment of the number of training data.

Domain	Way	Lexical tones	acc. (%)
Chinese Character uni-gram Pinyin			
News genre	2-way	Y	**85.94**
		N	85.52
	7-way	Y	**73.44**
		N	72.91
	20-way	Y	**78.64**
		N	75.19
Other genres	2-way	Y	**74.04**
		N	71.61
	7-way	Y	**68.96**
		N	66.99
	20-way	Y	**69.37**
		N	68.22
Chinese Character bi-gram Pinyin			
News genre	2-way	Y	63.56
		N	68.93
	7-way	Y	50.19
		N	53.43
	20-way	Y	50.59
		N	52.63
Other genres	2-way	Y	47.40
		N	52.89
	7-way	Y	43.39
		N	46.46
	20-way	Y	40.57
		N	45.08

Table 4: The performance of sentence-level Gan dialects identification using Chinese character Pinyin. 'Y' stands for the corpus with lexical tones, 'N' indicates none.

In addition, lexical tones in uni-gram Pinyin reflect the fine-grained characteristic of Gan dialects. Using lexical

tones is better than without it. We also conduct the tri-gram case, but the performance is lower than bi-gram about 20%. Compared with uni-gram, there are much noise in both bi-gram and tri-gram features. The proposed uni-gram features significantly outperforms the bi-gram ones with $p<0.05$ using paired t-test for significance. It shows the effectiveness of the proposed Chinese character Pinyin feature.

More specifically, the accuracy for each level-1 Gan dialects for news domain with Chinese character uni-gram Pinyin feature is reported in Table 5. As shown, we gain the best identification performance for 普通话 'Putonghua', while the accuracy of 鹰弋片 'yin yi region' is the worst one. The reason is that the difference between 鹰弋片 'yin yi region' and 普通话 'Putonghua' is not obvious enough as shown in Table 6, also we have enough training data for 普通话 'Putonghua'.

Dialect	acc. (%)
昌靖片 chang jing region	68.57
抚广片 fu guang region	76.09
客家话 Hakka	67.24
吉莲片 ji lian region	80.56
普通话 Putonghua	**95.50**
宜萍片 yi ping region	84.85
鹰弋片 yin yi region	22.73

Table 5: Accuracy of each level-1 Gan dialects on news domain.

To be more specific, we report the confusion table for each level-1 Gan dialect using Chinese character uni-gram Pinyin in Table 6. As can be seen, most instances have been correctly classified. Due to the challenge of discrimination for the closely related languages in the Gan dialects, some instances still have been falsely classified. For example, we can know that some instances falsely classified from 昌靖片'chang jing region' to'普通话' Putonghua (20) is similar to those from 鹰弋片'yin yi region' to 普通话 'Putonghua' (15). The reason is that the 昌靖片 'chang jing region' and 鹰弋片 'yin yi region' are closed to 普通话 'Putonghua'.

		Predicted label						
		L_1	L_2	L_3	L_4	L_5	L_6	L_7
True label	L_1	**48**	2	0	0	20	0	0
	L_2	2	**35**	2	1	6	0	0
	L_3	0	0	**39**	0	19	0	0
	L_4	1	1	4	**58**	8	0	0
	L_5	1	2	2	4	**212**	0	1
	L_6	1	0	1	0	3	**28**	0
	L_7	1	0	1	0	**15**	0	5

Table 6: Confusion table of each level-1 Gan dialects using uni-gram Pinyin with tones on news domain. Remark: L_1 stands for 昌靖片'chang jing region', L_2 indicates 抚广片'fu guang region', L_3 donates 客家话 'Hakka', L_4 refers to 吉莲片'ji lian region', L_5 means 普通话 'Putonghua', L_6 represents 宜萍片'yi ping region', L_7 embodies 鹰弋片'yin yi region'.

5.3.2 Results on Chinese Character

Table 7 shows the performance on Chinese character. Again, the character uni-gram feature yields best performance on both news and other type of genres. It also yields promising results on the extremely difficult fine-grained 20-way language classification.

Feature	Domain	Way	acc. (%)
Character uni-gram	News genre	2-way	**89.43**
		7-way	**76.03**
		20-way	**79.70**
	Other genres	2-way	**79.08**
		7-way	**67.70**
		20-way	**67.42**
Character bi-gram	News genre	2-way	73.27
		7-way	52.66
		20-way	52.02
	Other genres	2-way	58.71
		7-way	45.91
		20-way	44.35

Table 7: The performance of sentence-level Gan dialects identification using Chinese character.

6. Conclusions

In this paper, we annotate a parallel Gan Chinese Dialects Corpus (GCDC) based on different levels of modularity (written and spoken data) with different layers of annotations and transcription. Meanwhile, we conduct a preliminary experiment on the proposed GCDC through sentence-level Gan Chinese dialects identification task on different levels of granularity. The simple but effective character Chinese Pinyin and character uni-gram yields a strong baseline, especially on the 20-way Gan dialects discrimination, which shows the fine-grained automatic Gan Chinese dialects identification should be feasible.

In future work, we would like to explore more features without the need of using the Pinyin notation, enlarge the scale of the corpus, and test other classifiers. Furthermore, we will finally investigate how dialect identification can help other NLP tasks.

7. Acknowledgments

The authors would like to thank the anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant No. 61772246, 61402208, 61462044, 61462045, 61662031 and the Social Science Planning Project of Jiangxi Province under Grant No. 17YY05 and the humanities and Social Sciences projects of the Jiangxi Provincial Education Department under Grant No. YY17211.

8. Bibliographical References

Nancy F. Chen, Darren Wee, Rong Tong, Bin Ma, Haizhou Li. (216). Large-scale characterization of non-native mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. *Speech Communication*, 84:46-56.

Cagri Coltekin, taraka Rama. (2016). Discriminating similar languages : experiments with linear SVMs and neural networks. In *Proceedings of the third workshop on NLP for similar languages, varieties and dialects*, pages 15-24, Osaka, Japan, December 2016.

Heba Elfardy and Mona T Diab. (2013). Sentence level dialect identification in Arabic. *In Proceedings of ACL*, pages 456-461.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xang-Rui Wang, and Chih-Jen Lin. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871-1874.

Necip Fazil Ayan and Bonnie J Dorr. (2006). Going beyond aer: An extensive analysis of word alignments and their impact on MT. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual Meeting of the Association for Computational Linguistics*, pages 9-16.

Gregory Grefenstette. (1995). Comparing two language identification schemes. In *Proceedings of Analisi Statistica dei Dati Testuali (JADT)*, pages 263-268.

Nikola Ljubesic and Denis Kranjcic. (2015). Discriminating between closely related languages on twitter. *Informatica*, 39(1):1-8.

Marco Lui and Paul Cook. (2013). Classifying English documents by national dialect. In *Proceedings of Australasian Language Technology Workshop*, pages 5-15.

Bali Ranaivo-Malancon. (2006). Automatic identification of close languages-case study: Maly and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126-134.

Wolfgang Maier and Carlos Gomez-Rodriguez. (2014). Language variety identification in Spanish tweets. In *Proceedings of the LT4CloseLang Workshop*, pages 25-35.

Shervin Malmasi and Mark Dras. (2015). Automatic Language Identification for Persian and Dari texts. In *Proceedings of PACLING*, pages 59-64.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubesic, Preslav Nakov, Ahmed Ali, Jorg Tiedemann. (2016). Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of the COLING VarDial Workshop*, pages 1-14.

Rada Mihalcea and Ted Pedersen. (2003). An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1-10.

Kavi Narayana Murthy and G Bharadwaja Kumar. (2006). Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(01):57-80.

Matthew Purver. (2015). A Simple Baseline for Discriminating Similar Languages. In *Proceedings of the LT4VarDial*, pages 1-5.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. (2014). Sentence level dialect identification for machine translation system selection. In *Proceedings of ACL*, pages 772-778.

Alberto Simoes, Jose Joao Almeida, and Simon D Byers. (2014). Language identification: a neural network approach. In *Proceedings of the 3rd Symposium on Languages, Applications and Technologies, SLATE'14*, pages 252-265.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. (2002). Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 147-152.

Jorg Tiedemann and Nikola Ljubesic. (2012). Efficient discrimination between closely related languages. In *Proceedings of COLING*, pages 2619-2634.

Christoph Tillmann,Yaser Al-Onaizan, and Saab Mansour. (2014). Improved Sentence-Level Arabic Dialect Classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages*, pages 110–119.

Fan Xu, Xiongfei Xu, Mingwen Wang, Maoxi Li. (2015). Building Monolingual Word Alignment Corpus for the Greater China Region. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages*,Varieties and Dialects, pages 85–94, Hissar, Bulgaria, September 10, 2015. Association for Computational Linguistics.

Sen Yan. The Partitions of Jiangxi Dialects. *Dialect*, 1986,(1):19-38 (in Chinese).

Omar F Zaidan and Chris Callison-Burch. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1): 171-202.

Marcos Zampieri and Binyam Gebrekidan Gebre. (2012). Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS*, pages 233-237.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubesic, Preslav Nakov, Ahmed Ali, Jorg Tiedemann, Yves Scherrer, Noemi Aepli. (2017). Findings of the VarDial Evaluation Compaign. In *Proceedings of the EACL VarDial workshop*, pages 1-15.

Marcos Zampieri, Liling Tan, Nikola Ljubesic, and Jorg Tiedemann. (2014). A report on the DSL shared task 2014. In *Proceedings of the COLING VarDial Workshop*, pages 58-67.

Marcos Zampieri, Liling Tan, Nikola Ljubesic, Jorg Tiedemann, and Preslav Nakov. (2015). Overview of the DSL shared task 2015. In *Proceedings of the RANLP LT4VarDial workshop*, pages 1-9.