Language Identification for Person Names Based on Statistical Information

Shiho Nobesawa
Department of Information Sciences
Tokyo University of Science
2641 Yamazaki, Noda
Chiba, 278-8510, Japan
shiho@is.noda.tus.ac.jp

Ikuo Tahara
Department of Information Sciences
Tokyo University of Science
2641 Yamazaki, Noda
Chiba, 278-8510, Japan
Tahara@is.noda.tus.ac.jp

Abstract

Language identification has been an interesting and fascinating issue in natural language processing for decades, and there have been many researches on it. However, most of the researches are for documents, and though the possibility of high accuracy for shorter strings of characters, language identification for words or phrases has not been discussed much. In this paper we propose a statistical method of language identification for phrases, and show the empirical results for person names of 9 languages (12 areas). Our simple method based on *n*-gram and phrase length obtained more than 90% of accuracy for Japanese, Korean and Russian, and fair results for other languages except English. This result indicated the possibility of language identification for person names based on statistics, which is useful in multi-language person name detection and also let us expect the possibility of language identification for phrases with simple statistics-based methods.

1. Introduction

The technology of language identification has become more important with the growth of the WWW. As Grefenstette reported in their paper (Grefenstette 2000) non-English languages are growing in recent years on the WWW, and the need for automatic language identification for both documents and phrases are increasing. An easy method for language identification must be a key for better accuracy rate in natural language processing, such as information retrieval and machine translation.

Language identification is not a new topic in natural language processing. Bastrup proposed a unigram-based decision-tree method for language identification (Bastrup 2003). Dunning reported that 20 bytes are enough to obtain 92% accuracy in language identification (Dunning 1994). His method was based on statistical information, and it did not use any accented characters which would be a great help in identifying. This result is very encouraging for applying statistical language identification to proper nouns. Language identification is also well examined as speech recognition task (Matrouf 1998, Schultz 1996, Hazen 1994a, Hazen 1994b, Lamel 1994, Berkling 1994). Caseiro and Trancoso introduced a method using one language phone recognizer and less linguistic information for the language identification of speech (Caseiro 1998).

2. Language Identification for Person Names

Person name is one of the most frequent foreign phrases in texts, and it often causes noise as unknown words. Thus language identification of person names can be a help for better accuracy in text processing. However, there are several difficulties in identifying the language of a person name. First, the language of a person name may not match the official language of the area. Second, person names are international and not as language-specific as other words. Third, person names are often not long enough for analysis, and it is even possible to have name words in different languages in a full name. There can be found names in non-official languages of the area for many reasons such as international naming, international marriage or migration. And it is also difficult to identify a single language to a name, as there are common names used in several languages. So it is not practical to identify the language of a person name, however, to identify the area to where a person name belong should be possible with statistical data.

3. Statistics in Person Names

It is well known that each language has its own n-gram frequency (Dunning 1994). Our person name list corpora also have language-specific n-gram frequency in person names.

Thus we examined the possibility of language identification for person names based only on automatically-extractable statistical information. Nobesawa et al. proposed methods on obtaining domain-specific phrases using automatically-extractable statistical information only (Nobesawa 2002, Nobesawa 2000). Thus we examined the possibility of language identification for person names based only on automatically-extractable statistical information.

3.1. Corpora

We use person name lists for both training corpora and test corpora.

Our system requires statistical data extracted automatically from training corpora, and estimates the likelihood of each name belonging to each language.

3.1.1. Name Corpora by Areas

We prepared person name lists in 9 languages and made 12 person name corpora by areas (English (United Kingdom, United States), Chinese (mainland China, Hong Kong, and Taiwan), German, French, Greek, Japanese, Korean, Russian, and Thai.

As for person names, it is almost impossible to decide a single nationality to each name. Thus we work on area identification rather than language identification for person names, according to sets of names found in academic organization websites on the WWW. We can not avoid foreign names being noise with these data.

The size of the corpora varies, from at most 18,119 full names for Germany corpus to at least 5,764 full names for Greece corpora. Average number of full names in a corpus is 10,903 (Table 1). The basic data of the name corpora are shown in Table 1. “Name” in Table 1 means full names, “word” is for name words such as a single family name or a single given name, including initials. Table 1 also includes average number of words in a full name and average number of characters in a name word in each name corpus.

Table 1: Name Corpora for Experiments on Language Identification

name corpus	#name	#word	#char	#word/#name	#char/#word
China	7,388	15,004	86,336	2.03	5.75
Taiwan	7,676	16,486	105,010	2.15	6.37
Hong Kong	9,049	25,389	122,244	2.81	4.81
Korea	9,284	21,258	118,338	2.29	5.57
Japan	13,680	27,339	207,618	2.00	7.59
Thailand	6,774	13,693	129,415	2.02	9.45
Russia	5,891	14,564	122,663	2.47	8.42
Greece	5,764	11,749	100,705	2.04	8.57
France	14,405	30,005	232,667	2.08	7.75
Germany	18,119	37,406	277,188	2.06	7.41
U.K.	17,149	39,142	254,194	2.28	6.49
U.S.A.	15,661	34,212	231,259	2.18	6.76
average	10,903	23,853	165,636	2.20	7.08

3.1.2. Notation

Our system does not use the advantage of language-specified accented characters. We excluded names with accented characters from the corpora. Thus the corpora do not include, for example, “Uta Müller,” but they may include “Uta Muller” and “Uta Mueller.”

The corpora contain names with initials, such as “A.B. Chan” and “Ann B. Chan.” This has influence on the numbers per word/name in Table 1 and Figure 2, for the ratio of names with initials differ between the corpora. Some areas such as Hong Kong and Russia show tendency to have more initials.

3.2. Statistics in Person Names

Our system uses the statistical data of person names in the name corpora in Section 3.1.

3.2.1. Length

Names have small differences in their names according to the areas. The overall average number of characters in a full name is 16.19. Chinese and Korean is shorter in average, and Russian and Thai are longer.

Full-Name Length (Number of Words in a Full-Name) A Hong Kong name has 2.81 words on average, which is a unique feature (Table 1). A Chinese name is basically made of three characters, one for family name and succeeding two for a given name. But the basic notations of the names in Latin alphabet are different in China, Hong Kong and Taiwan. So-called English names are more common in Hong Kong, and many people put English names to/instead of Chinese given names. Thus there is variety of notation of names in Hong Kong. In Korea, person names are made of three words like in Chinese, and some people separate their given names in two words when writing in Latin alphabet. This is the main reason of having name-length average rather longer in Hong Kong and Korea comparing to other Asian areas.

In Russia the most frequent full-name length is three, where two is common in most of languages. Russian names have longer words than other languages, and this length information helps in distinguishing Russian from others, as Thai and Greek which also have longer words contain only two words in a name.

Word Length (Number of Characters in a Name Word) The average number of characters in a name word was 7.08 and the average number of name words in a full name was 2.20.

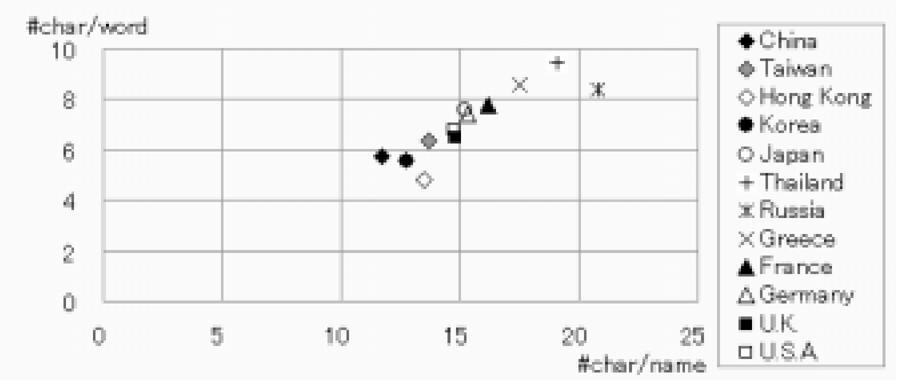

Figure 1: Average number of characters per word/full name of each corpus

As for the word length, Russian, Thai and Greece have slightly longer words. Figure 1 shows that there can be three groups according to the length; shorter-word group for 3 Chinese Areas and Korea, middle-length group for Europe, U.S.A. and Japan, and longer-word group for Thailand, Greece and Russia. Shorter-word group is tending to have approximately 6 characters. Japanese seems to be unique in Asian languages, as it has 8 letters in a word on average, which is more alike to middle-length group (Figure 2).

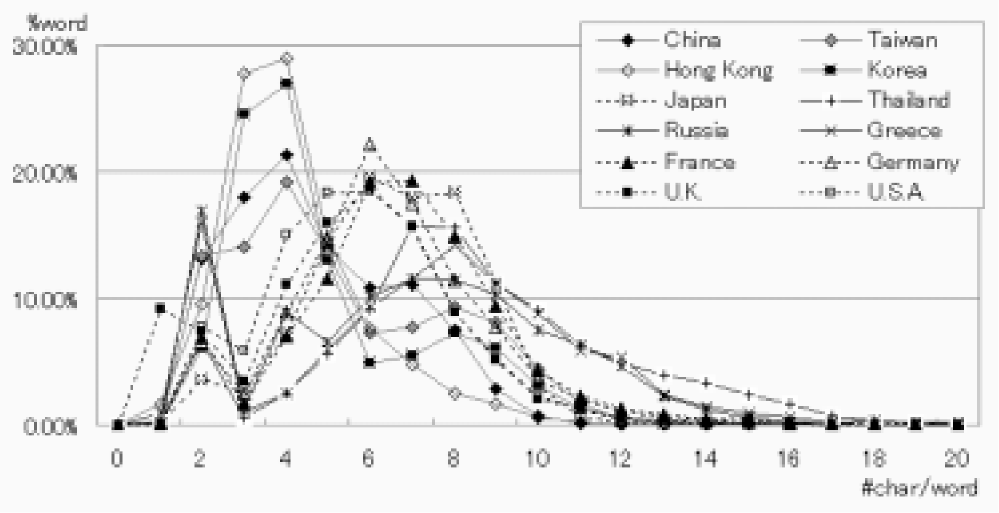

Figure 2: Ratio of Characters per Word (%)

3.2.2. *N*-gram

Unigram Table 2 shows the frequency of characters in each name corpus. Upper-case characters and lower-case characters are counted as the same.

Table 2: Unigram Ratio Ranking of Each Corpus

corpus	10% & more / 5% & more / 1% & more / less than 1%
China	N **I** G / **A** **U** H **E** / **O** Z Y L X J W S C Q D M / F T P B R K V
Taiwan	N H / **I** G **A** **E** **U** C / S L **O** Y W T J M R K / P F Z D B X V Q
H.K.	N / **A** **I** G H **E** **U** **O** / C L K W Y S M T R P / F D J Z B V X Q
Korea	N **O** / G **E** H **U** K **A** **I** / Y S M J L C W R P / B D T Z V X F Q
Japan	**A** **I** / **O** K H S **U** T / M R N Y **E** D G C W J / Z B F L P V X Q
Thailand	**A** N / T **I** R **O** H S / P **U** **E** K C M G W L D Y J / V B F Z X Q
Russia	**A** / **I** V **O** N **E** L R / K S H T M D Y G C **U** B P / Z F X J W Q
Greece	**A** **I** **O** S | N T R / L **E** K **U** P D G M H V C / Z F Y B J X W Q
France	**E** **A** / **I** R N L **O** / S T **U** C D H M B G P V F Y J / K Z Q W X
Germany	**E** / **A** R N **I** S L H T / **O** M C **U** K D G B F P W J / Z V Y X Q
U.K.	**A** / **E** R N **I** **O** L S / T H D M C G B **U** P W Y K J F V / Z X Q
U.S.A.	**A** **E** / N R **I** L **O** S / H T M D C G **U** K Y B J P W V F / Z X Q

As shown in Table 2, the ratio of characters differs with the areas.

Basic five vowels are emphasized in Table 2. ‘A’ and ‘I’ are frequent in almost all the corpora. However, for the Korea corpus ‘A’ and ‘I’ are the least frequent vowels, which is very different from other corpora. ‘U’ has high frequency in Asian corpora, but not in European corpora. ‘E’ is also frequent in most corpora, but the frequency rate differ comparing European corpora and Asian corpora.

‘Z’ and ‘Q’ are almost not used in 11 corpora, but are found in the China corpus, which is even different from the Hong Kong corpus and the Taiwan corpus. ‘V’ is also almost not used, but it is the most frequent consonant in the Russia corpus. ‘K’ is not a frequent character in most of the corpora, but it is the most frequent consonant in the Japan corpus. ‘T’ has low frequency ratio in Chinese corpora and the Korea corpus, but the ratio is more than 5% in Japan, Thailand, Greece and Germany.

Bigram There are language-specific features in bigram data as well. Bigrams mostly show the features of areas in vowel cooccurrences and consonant cooccurrences. Sometimes they are more area-specific rather than language-specific.

Trigram Trigram provides the specific features of the areas. The frequency rankings of trigrams differ between the areas, and high-ranked trigrams are the keys of high accuracy rate in area identification.

4. Identification Based on Statistical Information

Our system is fully based on statistical information. The system takes one full name as an input and outputs the possibility value for the input full name belonging to each name corpus.

4.1. Corpora

We used the name corpora explained in Section 3.1 as the test corpora for the experiments. Though the experiments were on closed corpora, the system may not be able to obtain 100% accuracy because of the foreign names included in the corpora.

4.2. Language Identification Methods Based on Statistical Data

The algorithm is the same for all the methods. The system calculates the possibility to belong to an area using the statistical data extracted from the name corpus. The estimation is calculated independently for each name corpus, thus the summation of the possibility ratio for each area is not 100%. This is because one name may have more than one area to belong, or no area, according to the listing up of the test areas. After examining all the areas, the system outputs the best-scored area as the result of the area identification of the input full name.

A Method Based on Length The system uses full-name length and word length as statistical data for estimation. This is rather a negative method to filter out areas with low possibility.

A Method Based on *N*-gram The system uses *n*-gram data for estimation. We had four patterns: unigram only, bigram only, trigram only, and interpolated trigram. Interpolated trigram is to avoid the problem of sparseness with trigram, and the system introduces bigrams for the gap. If the system finds lack of both trigram and bigram for a string, then unigram is used for interpolation. For bigrams and unigrams, the possibility value is reduced with certain weights.

A Method Based on *N*-gram and Length The system uses both interpolated trigram and length data for estimation. The weights for the two data should be controlled, and for the experiments in this paper interpolated trigram is more weighted so that length data is used as a support to raise the accuracy.

4.3. Empirical Results

We had experiments on area identification for the 12 name corpora. Table 3 shows the ratio of full names successfully estimated with our statistical methods. The results showed at most around 90% accuracy for four corpora (Japan, Russia, Korea and Thailand). Another four corpora showed 70% to 80% accuracy (Greece, Taiwan, Germany and China), and two corpora showed more than 60% accuracy (France and Hong Kong). The accuracy rate for U.K. was 59.18% and 50.88% for U.S.A. The accuracy rates were mostly good for most areas except English-speaking areas.

Table 3 indicated that interpolated *n*-gram was efficient for most of the corpora, and the system raised its accuracy rate slightly by combining length data and *n*-gram data. Length data were not efficient enough for an independent use, but they never decreased the accuracy when used combined with *n*-gram data.

From the results we recognized two groups of areas; the European-American group and the Chinese group. The European-American group includes U.K., U.S.A., France and Germany. Chinese group includes China, Taiwan and Hong Kong. The name corpora which belong to these groups showed lower accuracy rates, as they are often confused and mistaken between the areas in the same group. Obviously it is natural to have such area groups, and it can be said that we succeeded in recognition of area groups only with statistical information. Name corpora not included in the two area groups showed higher accuracy rates.

Table 3: Accuracy Rates on Language (Area) Identification (%)

name corpus	length	unigram	bigram	trigram	interpolated trigram	interpolated trigram + length
China	45.06	71.57	71.61	65.10	71.38	73.58
Taiwan	11.71	68.64	78.53	73.19	79.26	80.16
Hong Kong	55.92	12.53	56.96	56.99	58.64	63.56
Korea	7.63	63.23	84.24	85.92	91.07	91.41
Japan	45.15	70.12	88.82	91.81	92.89	92.93
Thailand	46.73	40.10	80.18	85.64	88.94	89.43
Russia	69.64	78.44	89.58	85.75	90.99	91.45
Greece	30.00	74.27	72.68	78.57	81.38	82.48
France	3.72	21.62	51.47	65.76	69.36	69.82
Germany	18.74	62.04	52.02	71.46	74.80	75.17
U.K.	18.00	.48	49.54	54.51	58.62	59.18
U.S.A.	9.12	38.12	28.97	43.40	47.92	50.88
average	30.12	50.10	67.05	71.51	75.44	76.67

The results in Table 3 show that it is possible to distinguish person names according to the areas they belong to. This is an encouraging result for simple language identification for phrases.

4.4. The Method Based on Length

The method based on length data was not efficient enough for most of the corpora. The best accuracy rate was 69.64% for Russia corpora. China, Hong Kong, Japan and Thailand obtained around 50% accuracy with length data, but for European corpora it was almost impossible to distinguish the area lists only with this length data. This was because length data did not have enough information for positive identification, and the data could only find out names with irregular length.

4.5. The Methods Based on *N*-grams

Methods based on *n*-grams showed better accuracy rates comparing to the length data (Table 3).

The Method Based on Unigram Unigram is useful for filtering out the names with infrequent alphabets. Unlike bigram or trigram, unigram works rather negative at that point.

Unigram data showed difference in efficiency. The Russia corpus gained 78.44% accuracy rate only with character unigram data. Greece, China and Japan also gained accuracy rate more than 70%, and Taiwan, Korea and Germany gained more than 60%. On the other hand, the accuracy rate was only 0.48% for the U.K. corpus. About half of the names in the U.K. corpus were misjudged to belong to U.S.A., and 30% to Germany. The U.S.A. corpus obtained 38.12% accuracy, but 30% of U.S.A. names were also misjudged to be German. The U.K. corpus and the U.S.A. corpus had almost the same result with unigram data, which sounds natural. The Hong Kong corpus also showed low accuracy rate (12.53%) with unigram data. The mixture of Chinese names and English names in notation was the main reason for this low accuracy.

The Method Based on Bigram With bigram data, the accuracy rates obviously rose comparing to the experiment with unigram. Bigram data were efficient enough for seven name corpora to show more than 70% accuracy. The best was 89.58% for the Russia corpus. Hong Kong, France, Germany and U.K. showed accuracy around 50%, and U.S.A. could gain only 28.97% accuracy with bigram data. There were confusion in U.K., U.S.A., Germany and France.

The Method Based on Trigram Trigram data showed better accuracy rates than unigram and bigram in ten corpora. Trigram can evaluate the areas' specific sequences of characters better than unigram and bigram, but person names are made of short strings of characters and the influence of data sparseness caused noise. For China the accuracy rate according to trigram data got down to 65.10%, from 71.61% with bigram. Accuracy rate with trigram was even worse than the rate with unigram for the China corpus. The Russia corpus also showed lower accuracy rate with trigram (85.75%) comparing to the rate with bigram (89.58%). For the U.S.A. corpus the rate rose to 43.40% with trigram data, though there still was the confusion in the European-American corpora

The Method Based on Interpolated Trigram Eleven out of twelve corpora raised their accuracy rate by using interpolated trigram, comparing to plain *n*-gram data. For the China corpus, the accuracy rate with interpolated trigram was 71.38%, which was a bit lower than 71.61% with bigram, and even lower than 71.57% of unigram. But the accuracy rate with trigram for the China corpus was 65.10%. Using interpolated trigram, the system almost recovered the loss.

4.6. The Method Based on Interpolated Trigram and Length

The main estimation method in this paper was the combination of length data and *n*-gram data (Table 3). With this method we succeeded in obtaining the best accuracy rates for all the 12 name corpora, including the China corpus. Length data were not efficient enough to distinguish areas, but the data helped to raise the accuracy rates, by wiping out the full names with exceptional length.

Table 4 shows the ratio of areas in the output for each test corpora. The sum of the values in each row is 100%. In Table 4, 75.58% of names in the China corpus were estimated to belong to China, 16.51% to Taiwan, and 3.30% to Hong Kong. Table 4 shows that our system succeeded in distinction of the three Chinese-speaking areas, and also it indicates that the system identified 95.39% names of China corpus belong to Chinese-speaking areas.

Table 4: Language Identification According to Interpolated Trigram and Length Data (%)

corpus	cn	tw	hk	kr	jp	th	ru	gr	fr	de	uk	us
China	**75.58**	16.51	3.30	1.24	.15	.31	.46	.08	.42	.29	.41	3.26
Taiwan	5.62	**80.16**	3.66	1.23	.17	.10	.70	.04	1.16	.86	.99	5.31
H.K.	7.51	8.90	**63.56**	1.49	.18	.22	.37	.03	1.44	.78	3.84	11.69
Korea	.42	2.57	2.26	**91.41**	.08	.19	.16	.02	.16	.32	.43	1.96
Japan	.69	.54	.23	.49	**92.93**	.42	.45	.15	.72	.55	.66	2.16
Thailand	.15	.41	.21	.06	.96	**89.43**	.24	.24	1.12	1.17	1.86	4.16
Russian	.06	.07	.06	.00	.07	.21	**91.45**	.30	.99	1.65	.98	4.16
Greece	.02	.36	.00	.00	.09	.36	2.41	**82.48**	4.23	2.45	2.13	5.79
France	.67	.56	.23	.11	.33	.46	1.40	1.01	**69.82**	8.63	5.36	11.41
Germany	.51	.36	.13	.33	.15	.29	1.68	.50	5.82	**75.17**	4.80	10.28
U.K.	.64	.78	.82	.27	.21	.78	.85	.85	5.65	7.17	**59.18**	22.79
U.S.A.	2.06	3.12	1.47	.92	1.24	1.40	1.27	.73	7.04	9.44	20.43	**50.88**

On the other hand, the accuracy rates of the U.K. corpus and the U.S.A. corpus are low. The U.K. corpus and the U.S.A. corpus are both in English language, and the system could not distinguish them. 81.97% of the names in the U.K. corpus were judged to belong to U.K. or U.S.A., and 71.31% in U.S.A. corpora was judged to be English. The U.S.A. corpus showed difficulty in area identification, which is because it contains more foreign (non-English) names than other corpora. This also caused misjudges for other corpora to belong to U.S.A.

4.7. Person Name Area Groups

The results show that the corpora can be divided into area groups according to their person names.

China, Taiwan and Hong Kong make an area group according to the results such as Table 2. Those three corpora showed similarities both in length and in *n*-gram data, which is natural because of the language they share. However, it was possible to distinguish these three with high accuracy rates with our method. This shows that our method obtained and used the areas' features efficiently. France, Germany, U.K. and U.S.A. also make an area group. But for the English-speaking areas, area identification was not very successful.

Other five name corpora (Japan, Korea, Thailand, Russia, and Greece) could be distinguished from others with high accuracy rates. Our method based on both length data and *n*-gram data enabled to estimate the likelihood of belonging to the areas.

5. Conclusion

In this paper we proposed a method for identifying languages of person names based only on statistical information. We had experiments with name corpora of 12 areas (9 languages), and showed that it was possible to identify the areas of person names only with statistical data. As the best results, we succeeded to obtain accuracy rates over 90% for Korea, Japan and Russia corpora. And also, we confirmed that our system recognized area groups based on languages.

This paper showed the possibility of automatic language identification of person names in Latin alphabet based only on automatically-extractable statistical information. We expect that better results lead us to better language identification not only for person names but for other phrases.

6. References

Argamon, S., I. Dagan and Y. Krymolowski. 1998. A Memory-Based Approach to Learning Shallow Natural Language Patterns. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp.67-73.

Berkling, K.M. and E. Barnard. 1994. Language Identification of Six Languages Based on a Common Set of Broad Phonemes. *Proceedings of the 3rd International Conference on Spoken Language Processing*, pp.1891-1894.

Borthwick, A. 1999. A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese. *Proceedings of the IREX Workshop*.

Bastrup, S. and C. Pöpper. 2003. Language Detection Based on Unigram Analysis and Decision Trees. *Language Processing and Computational Linguistics*.

Church, K.W. and P. Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, pp.76-89.

Caseiro, D. and I. Trancoso. 1998. Language Identification Using Minimum Linguistic Information. *Proceedings of the 10th Portuguese Conference on Pattern Recognition*.

Cucerzan, S. and D. Yarowsky. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. *Proceedings of the Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.90-99.

Dunning, T. 1994. Statistical Identification of Language. *Computer Research Laboratory Technical Report MCCS-94-273*, New Mexico State University.

Grefenstette, G. and J. Nioche. 2000. Estimation of English and Non-English Language use on the WWW. *Proceedings of the Recherche d'Information Assistée per Ordinateur*, pp.237-246.

Hazen, T.J. and V.W. Zue. 1994. Automatic Language Identification Using a Segment-Based Approach. *Proceedings of the 3rd International Conference on Spoken Language Processing*, pp.1883-1886.

Hazen, T.J. and V.W. Zue. 1994. Recent Improvements in an Approach to Segment-Based Automatic Language Identification. *Proceedings of the 3rd International Conference on Spoken Language Processing*, pp.1883-1886.

Lamel, L.F. and J.-L. Gauvain. 1994. Language Identification Using Phone-Based Acoustic Likelihoods. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.1, pp.293-296.

Matrouf, D., M. Adda-Decker, L.F. Lamel and J.-L. Gauvain. 1998. Language Identification Using Phone-Based Incorporating Lexical Information. *Proceedings of the 5th International Conference on Spoken Language Processing*.

Nobesawa, S., H. Saito and M. Nakanishi. 2000. Automatic Semantic Sequence Extraction from Unrestricted Non-Tagged texts. *Proceedings of the 18th International Conference on Computational Linguistics*, pp.579-585.

Nobesawa, S., K. Sato and H. Saito. 2002. The Use of Domain-Specific Statistical Data for Japanese Morphological Analysis. *Journal of Natural Language Processing*, vol.9. no.3, pp.21-40, written in Japanese.

Schultz, T., I. Rogina and A. Waibel. 1996. LVCSR-Based Language Identification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.781-784.