Determining the Origin and Structure of Person Names

Yu Fu[1], Feiyu Xu[2], Hans Uszkoreit[2]

IBM Software Group[1]
Zhangjiang Hi-Tech Park, Pudong New District
Shanghai 201203, China

Language Technology Lab, DFKI GmbH[2]
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany

E-mail: fuyush@cn.ibm.com, feiyu@dfki.de, uszkoreit@dfki.de

Abstract

This paper presents a novel system HENNA (Hybrid Person Name Analyzer) for identifying language origin and analyzing linguistic structures of person names. We conduct ME-based classification methods for the language origin identification and achieve very promising performance. We will show that word-internal character sequences provide surprisingly strong evidence for predicting the language origin of person names. Our approach is context-, language- and domain-independent and can thus be easily adapted to person names in or from other languages. Furthermore, we provide a novel strategy to handle origin ambiguities or multiple origins in a name. HENNA also provides a person name parser for the analysis of linguistic and knowledge structures of person names. All the knowledge about a person name in HENNA is modelled in a person-name ontology, including relationships between language origins, linguistic features and grammars of person names of a specific language and interpretation of name elements. The approaches presented here are useful extensions of the named entity recognition task.

1. Introduction

Human readers can often correctly identify origins of person names mentioned in newspapers, even if they do not really know or speak the original language. There seem to be distinctive patterns in names for distinguishing origins. In the natural language application, the language origin recognition of person names is very crucial for applications such as speech synthesis, machine translation and information extraction. Given a string chain such as “Size”, which is part of a Chinese name mentioned in a English sentence, this string will be pronounced completely differently from the noun “size” for measurement in English, namely “Si-Ze” instead of “size”. Business intelligence or other intelligence applications can really take advantage of this task: monitoring business people from a certain country in a certain business sector.

In our approach, we assume that the named entity recognition is already achieved. Thus, the input of our system is a name. Our task is to tell the origin of the names. In this context, one of the biggest challenges for us is the ambiguity of name origins. Many names or name components are often shared by two or more language origins. Another problem is that different parts in a name stem from different origins, e.g., names of American Chinese such as “Mary Wang”.

HENNA system presented here is able to detect or guess the origin of names in monolingual texts of a particular language (such as English, Chinese) by employing ME-based classifier. Furthermore, HENNA can analyze tnames and reveal their syntactic and semantic structures. In our studies, we obtain important insights into the conventions of person names for each language origin. We construct an ontology containing all the relevant linguistic knowledge that can contribute directly to language origin identification and name parsing tasks. This knowledge is employed for the name structure analysis.

The remainder of this paper is organized as follows: Section 2 describes research related work. Section 3 presents the system design and architecture. In Section 4, we explain our experiments and evaluations. Section 5 gives a short conclusion.

2. Related Work

Most approaches to language identification are based on statistical techniques. Grefenstette (1995) proposes an approach using short common words to identify the language of documents. This approach works well if big amount of textual data is available. Its performance degrades significantly when input sentences contain only a few words. Therefore, his work cannot be adapted to single person name origin recognition. Vitale (1991), Llitjos (2002) and Fei et al. (2005) utilize letter n-gram language models for language identification. Their experiments show that character-level subsequences provide strong evidence for predicting the language origin of proper names and documents. Chen and Rosenfeld (1999) make use of a maximum entropy model as their classification method and integrate many useful linguistic features for classification. In order to deal with data sparseness, Chen and Rosenfeld (1999) employ a Gaussian prior to improve the performance of their classifier.

However, most of the above approaches to the language origin recognition of person names just simply classify a name into one possible category, namely, a language. There are no detailed analysis of the internal structures of the names and the origins of their name components. They neither reports on solutions to the ambiguity problem nor to the multiple origins in one name.

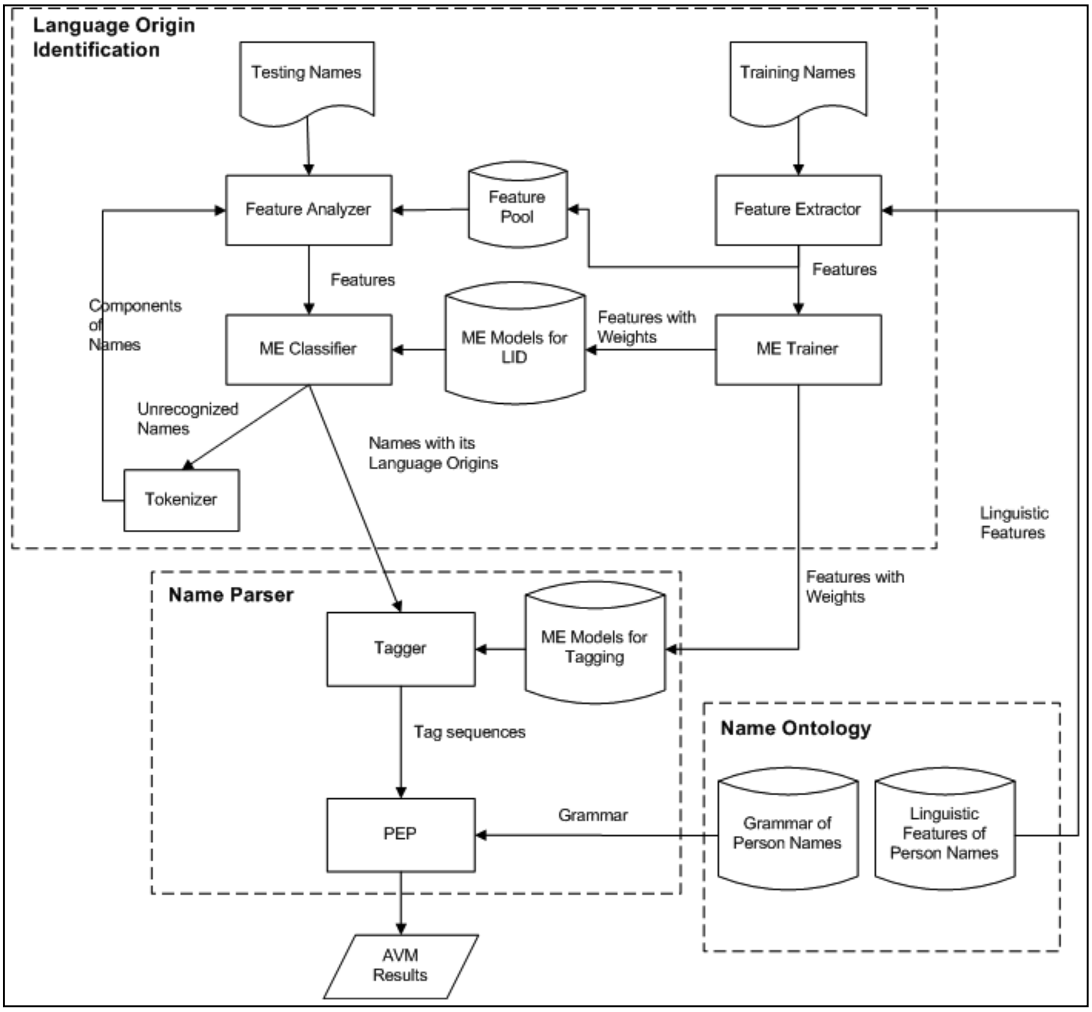


Figure 1: System Architecture of HENNA

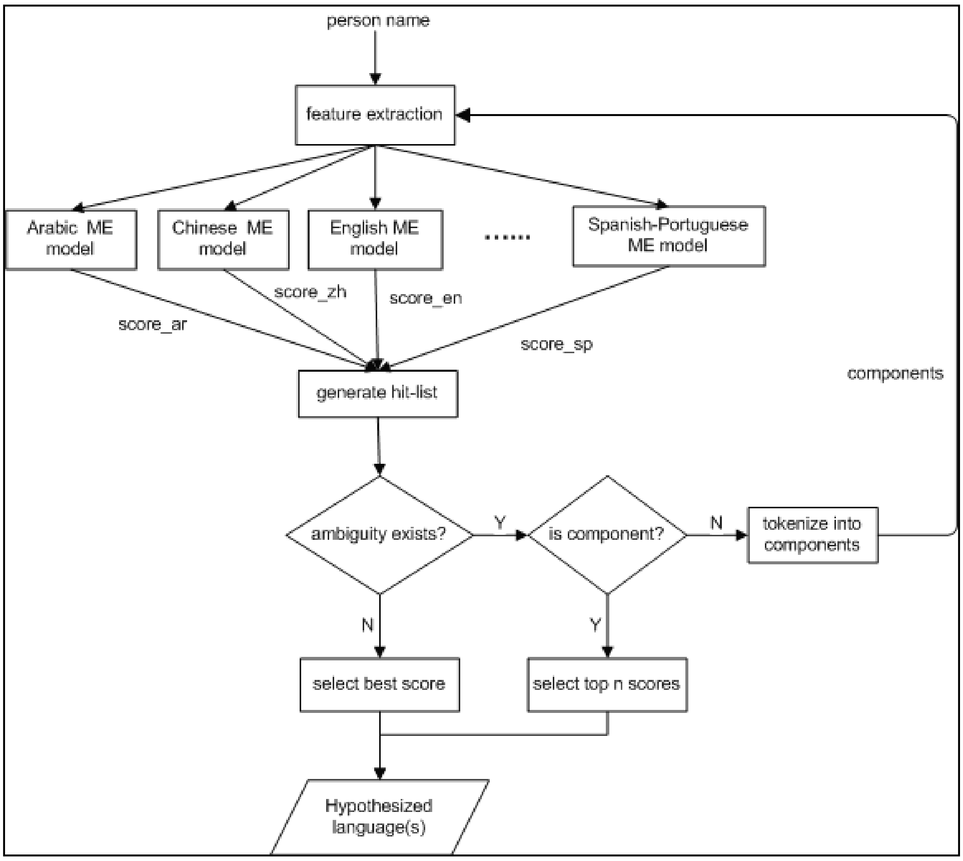


Figure 2: Workflow of the ME-based language identifier

Syntactic Rules	Example
PN → *KUNYA* + *ISM*	Abu Karim Muhammad
PN → *ISM* + *NASAB*	Yusuf ibn Ayyub
PN → *ISM* + *NISBA*	Muhammad al-Sakhtiyani
PN → *ISM* + *NASAB* + *NISBA*	Ahmad ibn Sa'id al-Bahili
PN → *KUNYA* + *ISM* + *NASAB* + *NISBA*	Abu Muqatil Muhammad ibn al-Munqadi al-Daylami

Table 1: Examples of Syntactic Rules of Arabic Names

Feature Type	Feature	Meaning	Example
kunya_preposition	*abu* *umm*	father of mother of	*Abu Karim Ahmad* *Umm Karim Ahmad*
nasab_preposition	*bin* *bint*	son of daughter of	*Yusuf bin Tariq* *Yusuf bint Ayyub*
prefix	*al*	the	*Muhammad al-Jamil*
common_name	*Muhammad*	praised	*Abu Karim Muhammad*

Table 2: Some Linguistic Features of Arabic Names

3. System Design and Architecture

Figure 1 gives a detailed view of the HENNA system. HENNA system contains two major subsystems: 1) language origin recognition; 2) name parser. Subsystem 1 contains a ME classifier incorporated with a rich feature set. A Markov classifier based on letter/character n-gram language model is also implemented in our experiments for comparison. Given the language origin of the names, Subsystem 2, the Person Name Parser, analyzes the linguistic structures of the person names. The results returned by the parser contain a structural analysis of the components in a name, as depicted in Figure 3. The name shown in Figure 3 has three components: given name, patronymic part and family name.

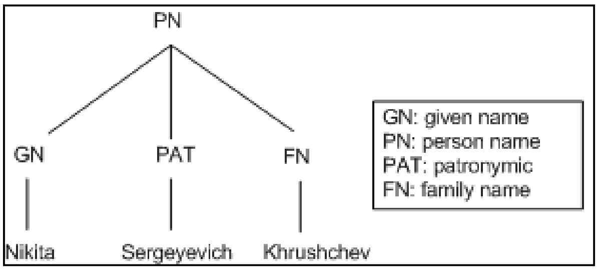


Figure 3: The Most Possible Parse Tree for Russian Name "*Nikita Sergeyevich Khrushchev*".

Figure 4 illustrates an attribute value structure (AVM) presentation of the analysis results of the above example name. It contains features about "surface string", "origin", "its name components" and "the interpretation of the patronymic part" and "its ranking about the analysis".

$$\left[\begin{array}{ll} \textbf{name} & \left[\ \textit{Nikita Sergeyevich Khrushchev}\ \right] \\ \textbf{origin} & \left[\ \textit{Russian}\ \right] \\ \textbf{syntax} & \left[\begin{array}{ll} \textbf{given name} & \textit{Nikita} \\ \textbf{family name} & \textit{Khrushchev} \\ \textbf{patronymic} & \textit{Sergeyevich} \end{array}\right] \\ \textbf{semantic} & \left[\ \textit{father's given name is Sergey}\ \right] \\ \textbf{credit} & \left[\ 0.9297\ \right] \end{array}\right]$$

Figure 4: The AVM Result for Russian Name "*Nikita Sergeyevich Khrushchev*".

Figure 2 gives a detailed description of the ME-based name origin recognition. For each language, a model is built. Given ambiguities of origins, names will be decomposed into their components, which will be classified to their origins separately. The advantage of this strategy is to allow multiple origins in a name. Furthermore, we still are able to classify our names into a language origin category based on their scoring.

4. Experiments and Results

Our name corpus is built on top of the following two major sources: 1) the LDC bilingual person name list and 2) the "*Person nach Staat*" (Person according to state) category of Wikipedia, which contains person names written in English texts from different countries.

Given a testing person name *W* labeled with language origin *L* in our corpus, if *L* has the maximal probability in the hit-list, name *W* is considered to be identified correctly in our experiments. The accuracy of language identification module of HENNA is evaluated by the equation shown below.

$$Accuracy = \frac{correctly\ identified\ person\ names}{all\ person\ names\ in\ testing\ corpus}$$

We use the results of the Markov-based language identifier as the baseline and compare them with the results of our ME-based language identifier. The comparison of the ME-based and the Markov-based classifier is shown in Table 3. Since Spanish and Portuguese are very closely related languages, and people in these countries share similar languages and cultural heritages, it's really hard to distinguish Spanish and Portuguese person name separately, so we decide to group Spanish and Portuguese into one language origin.

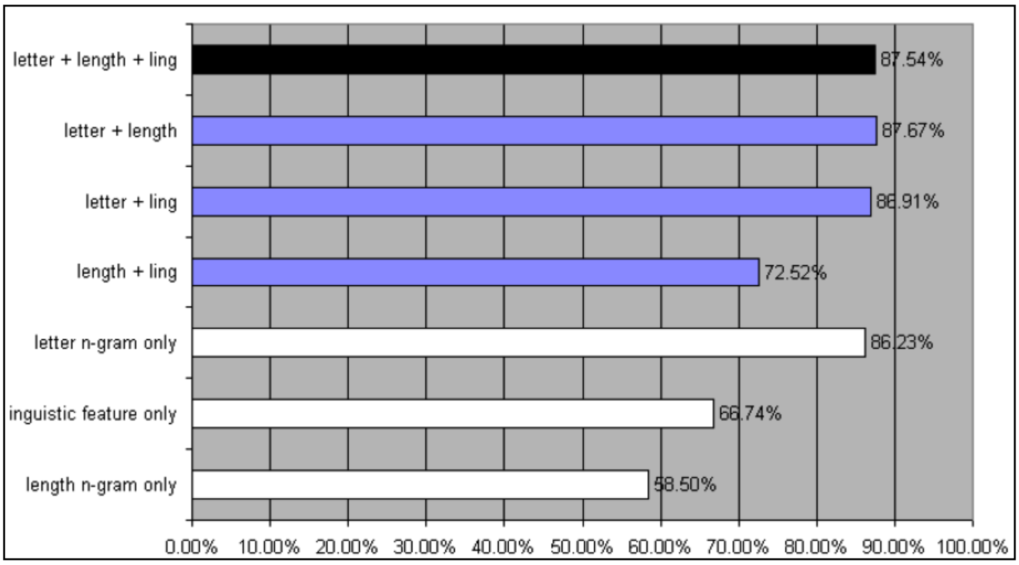
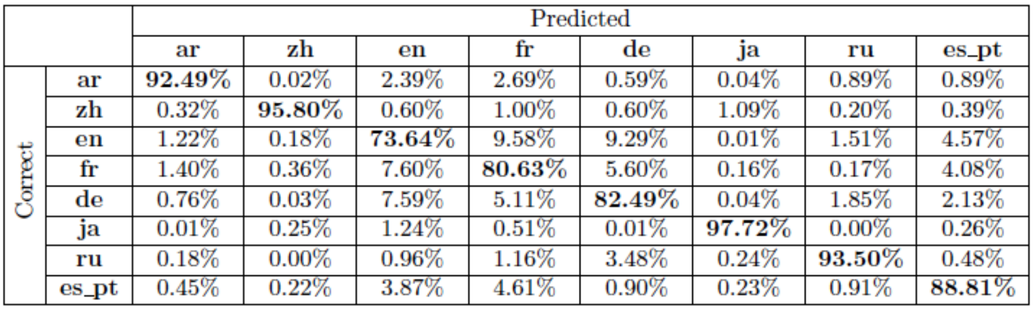
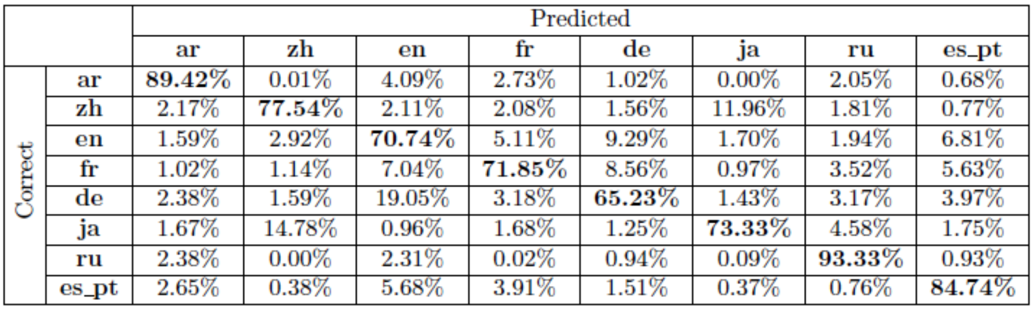


Figure 5: Classification Accuracy for Individual Feature Types and Their Combinations



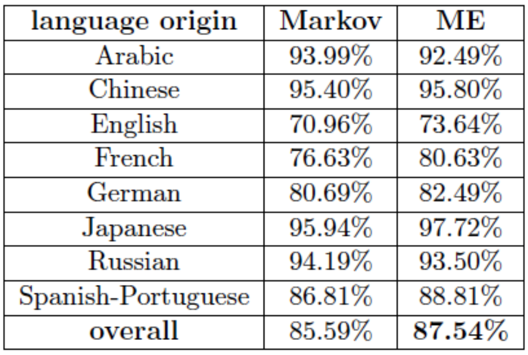
		Predicted							
		ar	**zh**	**en**	**fr**	**de**	**ja**	**ru**	**es_pt**
Correct	**ar**	**92.49%**	0.02%	2.39%	2.69%	0.59%	0.04%	0.89%	0.89%
	zh	0.32%	**95.80%**	0.60%	1.00%	0.60%	1.09%	0.20%	0.39%
	en	1.22%	0.18%	**73.64%**	9.58%	9.29%	0.01%	1.51%	4.57%
	fr	1.40%	0.36%	7.60%	**80.63%**	5.60%	0.16%	0.17%	4.08%
	de	0.76%	0.03%	7.59%	5.11%	**82.49%**	0.04%	1.85%	2.13%
	ja	0.01%	0.25%	1.24%	0.51%	0.01%	**97.72%**	0.00%	0.26%
	ru	0.18%	0.00%	0.96%	1.16%	3.48%	0.24%	**93.50%**	0.48%
	es_pt	0.45%	0.22%	3.87%	4.61%	0.90%	0.23%	0.91%	**88.81%**

Table 4.1: Confusion Matrix of N-way Test of Person Names in English Written Texts



		Predicted							
		ar	**zh**	**en**	**fr**	**de**	**ja**	**ru**	**es_pt**
Correct	**ar**	**89.42%**	0.01%	4.09%	2.73%	1.02%	0.00%	2.05%	0.68%
	zh	2.17%	**77.54%**	2.11%	2.08%	1.56%	11.96%	1.81%	0.77%
	en	1.59%	2.92%	**70.74%**	5.11%	9.29%	1.70%	1.94%	6.81%
	fr	1.02%	1.14%	7.04%	**71.85%**	8.56%	0.97%	3.52%	5.63%
	de	2.38%	1.59%	19.05%	3.18%	**65.23%**	1.43%	3.17%	3.97%
	ja	1.67%	14.78%	0.96%	1.68%	1.25%	**73.33%**	4.58%	1.75%
	ru	2.38%	0.00%	2.31%	0.02%	0.94%	0.09%	**93.33%**	0.93%
	es_pt	2.65%	0.38%	5.68%	3.91%	1.51%	0.37%	0.76%	**84.74%**

Table 4.2: Confusion Matrix of N-way Test of Person Names in Chinese Written Texts



language origin	Markov	ME
Arabic	93.99%	92.49%
Chinese	95.40%	95.80%
English	70.96%	73.64%
French	76.63%	80.63%
German	80.69%	82.49%
Japanese	95.94%	97.72%
Russian	94.19%	93.50%
Spanish-Portuguese	86.81%	88.81%
overall	85.59%	**87.54%**

Table 3: Comparison between ME-based Language Identifier and Markov-based Language Identifier

From Table 3, we can see that our ME-based identifier yields a similar or better performance than the Markov-based identifier for most languages, therefore, overall ME achieves an average performance of 87.54%, better than Markov with 85.59%. One potential reason may be that the ME model utilizes conditional likelihood as its natural objective function, which is normally more suitable for classification tasks than Markov models which are based on joint likelihood. Another likely reason is that more useful features can be integrated into the ME model, which however may lead to a heavy use of "back-off" in Markov modelling. Furthermore, the way in which different layers of back-offs are weighted against each other in a Markov model is also a weak point in comparison to the ME model. However, training an ME model usually requires much longer time than training a Markov model.

Three types of features are considered in the ME-based model of HENNA: letter n-grams, n-gram length and linguistic category. We test each feature type in isolation and each possible pair of feature types. The accuracy result is presented in Figure 5. We see that the letter n-gram feature is the most powerful single feature. The accuracies of feature combination reveal that n-gram length provides complementary information. Therefore, their combination wins against other features and their combinations. But combining linguistic category and the letter n-gram feature does not improve the performance. A possible reason may be that most linguistic category features such as prefix, suffix or common names are subsumed by letter n-gram features, these features could be eliminated without hurting classification performance.

Table 4.1 shows the confusion matrix of the n-way classification task. From this table, we can see that person names from English, German, French are most often confused, and some person names from French person names are often misclassified as Spanish-Portuguese person names. After we delete the French names, we rerun the n-way test and the average classification accuracy rise from 87.54% to 89.93%, and classification result of English person names is performed with an average accuracy of 80.64%. The difficulty of origin ambiguity has its historic reason, namely, the historical relationship between the languages. For instance, both English and German belong to the Germanic language family and thus many names are shared by both languages. The good performance of our language identification can partially be attributed to the distinctive spelling regularities of Arabic, Chinese and Japanese, which are quite different from other languages.

We also conduct experiments on person names in Chinese texts. Table 4.2 shows the confusion matrix of the n-way classification task. Comparing the figures with Table 4.1, we find that the performance achieved by classifying names in English texts is always better than the corresponding results in the task applying to Chinese texts. One reason may be that the English writing system has a limited set of 26-letter alphabets. With a small set of training data, our classifier can easily achieve its full coverage of the potential letter combination in English texts. However, this seems to be difficult for person names in Chinese written texts, since the Chinese writing system has more than ten thousands of characters or signs and the big potential of various combinations easily leads to a data sparseness problem and thus requires a heavy use of backing-off. Another likely reason for lower accuracies in classifying names from Chinese texts is the information loss due to transliteration of foreign names into Chinese signs.

However these linguistic features are very helpful in person name parsing, which enrich the linguistic expressiveness and contribute to the semantic interpretation of the name components.

5. Conclusion

This paper presents a system that identifies or guesses the language origin of person names and analyzes person names in English texts according to their linguistic structure. The system only uses name-internal features including surface string patterns and linguistic categories, automatically learned from data.

We have shown that word-internal character sequences provide surprisingly strong evidence for predicting the language origin of person names. Our approach is context-, language- and domain-independent and can thus be easily adapted to person names in or from other languages. It may also be applicable to classification tasks in other domains such as company or product names.

The experiments show that HENNA's language identification module works accurately and robustly for person names in English written texts. The language origin information and internal structure of a person name could be an important auxiliary input for many NLP applications, such as TTS, MT, QA or NER systems.

6. Acknowledgements

This work would not have been possible without the support of many people inside and outside of the Saarland University and the DFKI GmbH, we want to thank all the people that have educated, supported, and challenged us in Germany. The work presented here was partially supported by a grant to the project KomParse by the ProFIT programme of the Federal State of Berlin, co-financed by the European Union through the EFRE fund (Europäischer Fonds für regionale Entwicklung / European Regional Development Fund), as well as through a research grant to the project TAKE, funded by the German Ministry for Education and Research (BMBF, FKZ: 01IW08003).

7. References

Grefenstette, G. (1995). Comparing Two Language Identification Schemes. Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data.

Vitale, T. (1991). An Algorithm for High Accuracy Name Pronunciation by Parametric Speech Synthesizer. *Computational Linguistics,* 17(3), pp:257-276.

Berger, A.L., Della Pietra, V.J. , and Della Pietra, S.A. (1996). A Maximum Entropy Approach to Natural Language Processing. Capturing and interacting with design history. In *Computational Linguistic*s, 22(1), pp:39-71.

Borthwick, A. (1999). A Maximum Entropy Approach to Named Entity Recognition. *PhD thesis,* New York University.

Chen, S.F. and Maison, B. (2003). Using Place Name Data to Train Language Identification Models. *In Eighth European Conference on Speech Communication and Technology.*

Chen, S.F. and Rosenfeld, R. (1999). A Gaussian Prior for Smoothing Maximum Entropy Models. Techn. *Report CMU-CS-99-108,* Carnegie Mellon University.

Chen, Y., You, J., Chu, M., Zhao, Y., and Wang, J. (2006). Identifying Language Origin of Person Names with N-Grams of Different Units. *Acoustics, Speech and Signal Processing*,.

Church, K. (1985). Stress Assignment in Letter to Sound Rules for Speech Synthesis. *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pp : 246-253.

Lewis, S., McGrath, K., and Reuppel, J. (2004). Language Identification and Language Specific Letter-to-Sound Rules. *Colorado Research in Linguistics*, 17(1) , pp:1-8.

Llitjos, A.F. (2002). Improving Pronunciation Accuracy of Proper Names with Language Origin Classes. *Proceedings of the Seventh ESSLLI Student Session*, pp: 1-17.

Nigam, K., Lafferty, J., and McCallum, A. (1999). Using Maximum Entropy for Text Classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp: 61-67.

Fei, H., Stephan, V. and Alex, W. (2005). Clustering and Classifying Person Names by Origin. AAAI, pp: 1056-106