# Identifying Foreign Person Names in Chinese Text

**Stephan Busemann, Yajing Zhang**

DFKI GmbH
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
<firstname>.<lastname>@dfki.de

**Abstract**

Foreign name expressions written in Chinese characters are difficult to recognize since the sequence of characters represents the Chinese pronunciation of the name. This paper suggests that known English or German person names can reliably be identified on the basis of the similarity between the Chinese and the foreign pronunciation. In addition to locating a person name in the text and learning that it is foreign, the corresponding foreign name is identified, thus gaining precious additional information for cross-lingual applications. This idea is implemented as a statistical module into a rule-based named entity recognition system.

## 1. Introduction

The named entity recognition (NER) task for Chinese is particularly important because of the increasing number of online documents available in Chinese. The more these documents deal with international themes, the more frequently they contain transliterated foreign[1] words. The recognition of translated foreign names in Chinese plays an important role in areas such as cross-lingual information extraction and machine translation. Among the 2872 person names found in our newspaper corpus, 526 were foreign.

Foreign name expressions written in Chinese characters are difficult to recognize since the sequence of characters usually represents the Chinese pronunciation of the name.[2] Thus the length of a transliterated name is closely related to the number of syllables of the original foreign name. This makes the recognition task difficult as long as the original pronunciation is unknown.

In view of the lack of any electronically available standard repository mapping between foreign person names and a corresponding sequence of Chinese characters, we suggest that known foreign person names can reliably be identified automatically on the basis of the similarity between the Chinese and the foreign pronunciation. In addition to locating a person name in the text and learning that it is foreign, the corresponding original name is identified, thus gaining precious additional information for cross-lingual applications. This idea is implemented for German and English names as a statistical module into a rule-based named entity recognition system, using Mandarin as the Chinese source language.

More precisely, we combine rule-based NER with the statistical computation of the similarity between the Pinyin transcript of a foreign name candidate and the phonetic representation of pre-stored foreign names, accounting for the above-mentioned facts on foreign name encoding. This approach has to our knowledge not been investigated before. This paper shows that it is valid and provides promising results.

The reminder of the paper is organized as follows. The nature of the task is described in Section 2. Work closely related to the present task is sketched in Section 3. Section 4. describes how phonetic similarity is computed. The system implementation is presented in Section 5. followed by an evaluation and an error analysis in Section 6. Conclusions and further directions are given in Section 7.

## 2. On Transliteration

In the context identifying proper names, Lee et al. (2006) state that "transliteration is the process that converts an original proper name in the source language into an approximate phonetic equivalent in the target language, whereas back-transliteration is the reverse process that converts the transliteration back into its original proper name."

Transcoding the original pronunciation into Chinese characters is possible since the Chinese language has over 400 unique syllables (without counting tone variations), enough to approximate syllables which appear in other languages. One Chinese syllable can be represented by different characters. Therefore, there is a wide range of homonyms. For instance, the syllable *si* can be written as 丝 (meaning "silk"), 思 (meaning "thinking"), 死 (meaning "die") or 饲 (meaning "feed"), etc. However, not all these homonyms can be used for foreign name transliteration, as some of them have negative connotations, some are typical for Chinese surnames only, etc.

Selecting a character among all its homonyms may lead to different transliterations of the same foreign name. For instance, the former American president *Clinton* can be transliterated into 克林顿 (*ke4-lin2-dun4*) and 柯林顿 (*ke1-lin2-dun4*). The former version is predominantly used in Mainland China, whereas the latter is mostly found in Taiwan.

In recognizing foreign person names in written Chinese and considering a back-transliteration in Latin script, the following aspects also need consideration (cf. also (Knight and Graehl, 1998) for comparable issues in the back-transliteration from Japanese to English). Should, at word endings, the final consonant be left out, or be transliterated with a subsequent vowel? Consider

[1] By "foreign", we refer in the present paper to non-Chinese language.

[2] There are exceptions though, e.g. Japanese or Korean names as well as names in many languages spoken by minorities in China.

(1) Mubarak 穆巴拉克 (*mu4-ba1-la1-**ke4***)
穆巴拉 (*mu4-ba1-la1*)

Second, phonetic similarity may be judged differently, as with

(2) da Vinci 达芬奇(*da2-**fen1-qi2***)
达文西(*da2-**wen2-xi1***)

Third, the origin of a name plays an important role. For instance, the transliteration of the name *Jean* into Chinese may be based either on the French or the English pronunciation.

Finally, the same foreign name might be transliterated differently, depending on the use of the various dialects spoken in Mainland China, in Hong Kong, and in Taiwan, a theme addressed by Huang et al. (2007). Furthermore, Taiwan uses Traditional Chinese, whereas Simplified Chinese is the current standard writing system in Mainland China.

While we ignore, in the present work, the last point by simply sticking to Mandarin, we cover the first two aspects by designing suitable similarity metrics to compare pronunciations (see Section 4.). The third aspect is accounted for by assuming that phonetic representations of foreign names are available, which enables us to come up with an interpretation like *"Jean (English)"*. We currently account for German and English phonetic representations.

## 3. Related Work

As Chinese word segmentation is best based on semantic and contextual information, NER is often designed to be a byproduct of word segmentation.

Gao et al. (2004) see NEs as one out of five different classes Chinese words can be classified into. They use an n-gram language model for generic segmentation and transformation-based learning to adapt the output to different application-specific standards.

In (Zhang et al., 2003) a stochastic role model is constructed together with a lexicon, in which known Chinese NEs are listed. A corpus containing words and their POS tags is annotated automatically with role labels defined on the basis of linguistic features. Roles include surname, prefix to a name, tokens between two NEs, etc. The NER task is accomplished by searching the longest match on the best role sequence using the Viterbi algorithm.

A distinction between recognizing Chinese and foreign person names is made by Chen and Lee (1996). They rely on character frequency, assuming, as we do, that foreign person names are encoded using a set of a few hundred Chinese characters only. Their NER is corpus-based and uses character frequency information, which is different from our approach.

Lin and Chen (2002) propose a method for transliterating foreign names in Chinese into original English forms. They convert both the transliterated Chinese names and the original English names into phonemes of the International Phonetic Alphabet (IPA)[3] which are mapped onto the machine-readable SAMPA alphabet (Wells, 1997).[4] The English candidate with the highest similarity score is selected as the correct original word. The transformation of a Chinese name into the IPA representation is accomplished in two steps: converting a Chinese character into Pinyin symbols and mapping the initial consonant and the remaining vowel into IPA by looking up a table defined by Hieronyms (1997).

Lin's and Chen's work relies on input that is known to correspond to foreign names, whereas in the present work, a similar type of comparison is used to detect foreign person names, i.e. we have to deal with noisy data.

| Substitution | | 0.5 |
|---|---|---|
| Deletion | | 0.2 |
| Insertion | | 0.3 |
| Pinyin | SAMPA | Cost |
| te | t | 0.1 |
| si | s | 0.0 |
| l | r | 0.2 |
| a | @ | 0.0 |
| en | En | 0.0 |
| ang | {m | 0.0 |

Table 1: Excerpt of the Similarity Metric Used in SILO.

## 4. Computing Phonetic Similarity

Phonetic similarity is best computed using a common alphabet. Pinyin (Yin and Felley, 1990) was chosen as the standard phonetic transcription of Mandarin into Latin script. Transcribing Chinese characters into Pinyin was achieved with help of an open source Pinyin converter available from the Internet.[5] For the set of foreign names we chose a phonetic transcription (SAMPA) rather than their lexical representation. This way the comparison is more independent of the language according to which the name is pronounced.

The similarity of two strings (Pinyin, SAMPA) can be compared in various ways. One common approach is to calculate the edit distance of two strings using substitution, insertion and deletion operations as described in (Levenshtein, 1966).

We use the SILO tool (Eisele and vor der Brück, 2004), which allows the developer to define a metric of individual costs for insertion, deletion and substitution operations, with a default for operations not explicitly specified. SILO returns all entries in the lexicon whose costs with respect to an input word are below a given threshold.

In our system, the threshold was empirically set to 0.4, with the default costs for operations as given in Table 1. The table shows relations between Pinyin and SAMPA symbols whose costs differ from the default. For instance, each Chinese syllable ends in a vowel, which is not the case in German or English. To prevent the corresponding substitutions from being penalized, they must be explicitly listed. Figure 1 shows an example for the recognition of the English name *Sampras*, using the above excerpt of the metric. The figure shows an optimum alignment between Pinyin and

[3] http://www.arts.gla.ac.uk/IPA/ipachart.html

[4] http://www.phon.ucl.ac.uk/home/sampa/index.html

[5] Author: Jisheng Xie, http://okone96.itpub.net/post/9033/222538

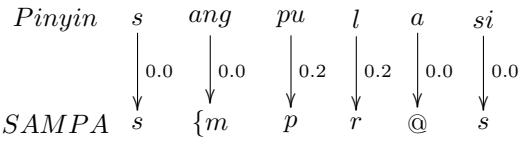


Figure 1: Optimum Alignment between *sangpulasi* and *s{mpr@s* in the Course of Transliteration of 桑普拉斯 into *Sampras*.

SAMPA at a total cost of 0.4. All other possible alignments would come at a higher cost.

## 5. The HyFex NER System

### 5.1. Architecture

Figure 2 gives an overview of HyFex (Hybrid Foreign Name Extraction) and its components. The system is based on the shallow parsing system SProUT (Shallow Processing with Unification and Typed Feature Structures; (Drozdzynski et al., 2004)), which offers rule sets and knowledge sources for NER in a multitude of languages. For the preprocessing of Chinese text, SProUT uses the tokenizer from the University of Shanxi, which also performs POS tagging (Liu, 2001). SProUT does not only deliver typed substrings, but structured information (see Figure 4). In addition to other NEs, SProUT recognizes Chinese and foreign person names stored in a gazetteer (Section 5.2.). SProUT interprets parts of a text using hand-written rules (Section 5.3.).

The statistical module, which is responsible for the similarity comparison of two pronunciations, is integrated into SProUT for recognizing the transliterated foreign names not covered by rules. This component is triggered by a sequence of special "trigger" characters (Section 5.4.) that frequently encode foreign names. The first step in the statistical module is the conversion of Chinese characters into Pinyin. It is followed by the similarity computation in SILO, as described in Section 4. SILO relies on a gazetteer associating names and their corresponding SAMPA representations. These phonetic name representations are computed offline with help of the text-to-speech system MARY (Schröder and Trouvain, 2003). Input to MARY is a list of pre-stored foreign names for either German or English.[6]

### 5.2. Gazetteers

SProUT gazetteer entries are not mere words, but rather words annotated by feature-value pairs representing type and other structural and semantic information. The gazetteer of Chinese and foreign names in the rule-based module (Figure 2) is restricted to a set of very frequent first names and last names (about 800 entries). Besides, it contains the trigger characters and some other information useful for disambiguation. Entries look as follows:

```
布什 | GTYPE: zh_person_name | LATIN: "Bush"
斯   | GTYPE: zh_trigger
经济学家 | GTYPE: zh_position
     | PROFESSION: "Economist"
```

The gazetteer of phonetic symbols is a mapping from the phonetic representations of person names, as they are generated by MARY, onto the respective SProUT gazetteer entries. It is used by the statistical module only. This way it is possible to retrieve any type of information encoded in the SProUT gazetteer, which may be used to facilitate disambiguation. Entries might look as follows:

```
  pIrs →   Pearce | LANGUAGE: EN | ...
  pIrs →   Peirce | LANGUAGE: EN | ...
 da:vit →  David | LANGUAGE: DE | ...
 dEIvid →  David | LANGUAGE: EN | ...
```

Gazetteers can be maintained independently. The phonetic gazetteer this system was tested with has around 81,000 English or German entries for person names.

### 5.3. SProUT

The rule-based part of HyFex is realized by SProUT. For extensive information about SProUT, we refer to (Drozdzynski et al., 2004). In the present context, we only highlight some relevant aspects of the system. In SProUT typed feature structures are used as a data structure to represent all linguistic objects (Krieger et al., 2004). SProUT rules combine the results of tokenization, morphological analysis and gazetteer-based analysis to recognize larger text segments and assign meaning to them. A rule matches the longest sequence of input tokens possible. The left hand side of a rule (all material preceding the arrow) makes use of regular expressions over typed feature structures to represent the recognition pattern. On the right-hand side, a typed feature structure represents the output data structure.

```
foreign_person :>
   gazetteer & [ GTYPE zh_person_position,
                 PROFESSION #position ]?
   gazetteer & [ GTYPE zh_person_name,
                 SURFACE #zh1, LATIN #n1 ]
   gazetteer & [ GTYPE zh_name_separator,
                 SURFACE #sep ]
   gazetteer & [ GTYPE zh_person_name,
                 SURFACE #zh2, LATIN #n2 ]
-> ne-person &
[SURFACE #surface, P-POSITION #position,
 GIVEN_NAME #n1, SURNAME #n2 ],
 where #surface = Append(#zh1, #sep, #zh2).
```

Figure 3: A SProUT Rule Named `foreign_person` That Matches Gazetteer Structures.

The sample rule in Figure 3 analyzes person names that occur in the gazetteer of foreign and Chinese names. The output is of type `ne_person`, which is structured into Latin given and surname, and position. The above rule accepts a string like 经济学家戴维．皮尔斯 (Economist David Pearce) that can be subdivided into three or four parts (the position part is optional), which in turn are defined as types

[6] Thus our system does not *guess* foreign names from pronunciations, which would be another direction of research.

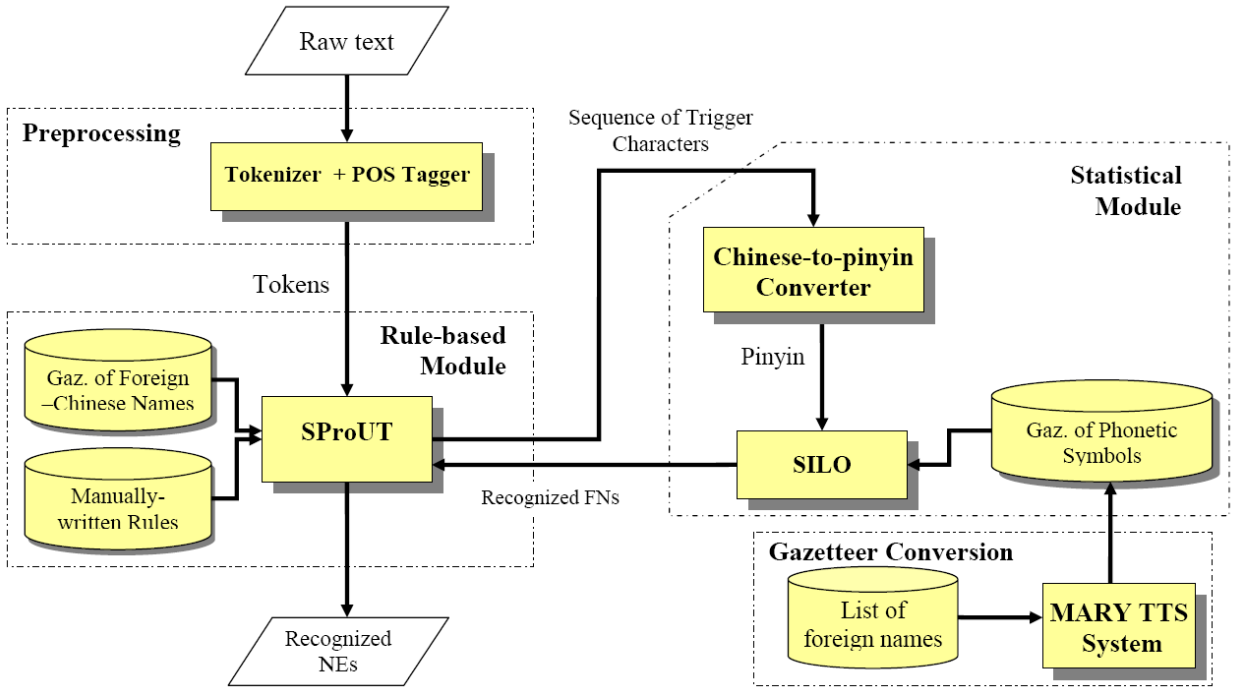


Figure 2: Data Flow Diagram of the HyFex System.

(all are required to be gazetteer entries). The relation between input and output is established using coreference of variables, which are prefixed by '#'. Note the use of the functional operator `Append`, which reconstructs the Chinese name.

A SProUT rule may fail if one of the matches defined on the left-hand side is not successful. It may also fail if a functional operator fails. Usage of `Append` in the sample rule always yields a value if the matches succeed and thus all parameters are bound.

SProUT output is displayed as a feature structure (cf. Figure 4), but also encoded in a corresponding XML format for further processing.

### 5.4. Trigger characters

Chinese characters used for foreign names are limited. As the sets used in related work were not available to us, we manually created a set as follows. 363 characters were extracted from about 700 person names in the gazetteer of foreign and Chinese names. To cover German names, characters used in the German Person Name Transliteration Manual (Xinhua News Agency, 1999) were also taken into account, yielding another 90.

Unfortunately this collection contained characters such as 二 (two) and 日 (day) in 阿巴斯二世 (*Abbas II*) and 拉格朗日 (*Lagrange*), respectively, which are highly ambiguous and not even representative for foreign names. They were removed from the set, sacrificing some recall (appx 2%) in favour of increased precision (appx 10%) of the overall system. We ended up with a set of 353 characters. This set can be modified and tuned to specific applications.

### 5.5. Implementation

We discuss now the processing according to Figure 2. A string of Chinese characters is first preprocessed using the Shanxi word segmenter. The result is subjected to SProUT rules that can identify pre-stored person names from the gazetteer of Chinese and foreign names and some unknown person names. There are SProUT rules that dispatch – upon recognizing a string of trigger characters – to the statistical module. The string is converted into Pinyin. The result is compared to the precomputed set of phonetic name representations using SILO. The best fit is returned, and a corresponding name (together with other gazetteer information) is returned to SProUT.

```
sprout_rule
NAME  foreign_person
OUT   ne-person
        CSTART      "1"
        CEND        "10"
        AGE         string
        P-POSITION  "Economist"
        TITLE       *opencons*
        SURFACE     "戴维·皮尔斯"
        SURNAME     "Pearce"
        GIVEN_NAME  "David"
```

Figure 4: A Sample Feature Structure for Output from the Rule in Figure 3. Note that the type `ne_person` induces more features than are used in the rule, i.e. `AGE`, `TITLE`, `CSTART`, `CEND`. The latter two are the character positions of the matched piece of text and automatically instantiated by the SProUT interpreter.

The statistical module is implemented as a functional operator *CombineStatistics* in SProUT, taking a sequence of Chinese trigger characters as input and returning, if successful, an English or German name. If no name is found below the threshold, the operator fails, causing the embedding SProUT rule to fail as well.
The length of the sequence of trigger characters may range between two and seven (foreign first or last names are rarely transliterated with more than seven characters). The trigger characters are stored under a dedicated type – `zh_trigger` – in a gazetteer.
The sample rule in Figure 5 collects six trigger characters and produces a feature structure of type `ne_person` with the name returned by the statistical module in Latin script.

```
foreign_person_stat :>
   gazetteer & [ GTYPE zh_trigger,
                 SURFACE %<char> ]{6}
-> ne-person & [SURFACE %<char>,
   SURNAME #sname, GIVEN_NAME #gname],
   where <#sname,#gname>
         = CombineStatistics(%<char>).
```

Figure 5: A SProUT Rule Interfacing the Rule-Based and the Statistical Modules (%<char> is a list variable.)

The current implementation returns only a single result. However, the SILO comparison can yield multiple results ranging below its threshold of error tolerance. Moreover, a result may correspond to multiple foreign names. For instance, *pIrs* may correspond to individuals named Pierce, Peirce, or Pearce. While in SProUT multiple results can be represented using feature structure lists, we must leave it to future work to augment the interface to the statistical module accordingly.
Moreover, by using position or profession information found in the text, we will be able to disambiguate between e.g. the mathematician David Pierce and the economist David Pearce.
By taking up suggestions of Li et al. (2007) for transliteration, gender information can be exploited for back-transliteration in a similar way.

## 6. Evaluation

### 6.1. Results

We based our formal evaluation on the publicly available pre-annotated January 1998 issues of People's Daily newspaper[7], which contains around 1.1 million words. The corpus has been hand-annotated at the Institute of Computational Linguistics in Peking University and the Fujitsu Research and Development Center. The texts cover various genres including politics, music, sports, poetry, etc. Most of the foreign person names contained can be attributed to politics and sports. They had to be tagged manually in addition.
The results achieved by a purely gazetteer-based SProUT version for Chinese NER, which also includes the Shanxi word segmenter, form a baseline, for which complete or partial identification are counted as true positives. This system corresponds to the left half of Figure 2.
As partial information may be valuable for many applications, we evaluated the hybrid system – i.e. the baseline system plus the statistical component – according to two principles:

**Exact** : Only the correctly recognized names with their correct transliterations are considered true positives.

**Indicative** : In addition, correctly recognized names with false transliterations, or partially recognized names (first names or last names), are counted as true positives.

Being gazetteer-based, the baseline system always produces the correct transliteration after completely identifying a name. In either case, Chinese person names recognized as foreign count as false positives.
HyFex has been trained on 5/6 of our People's Daily corpus and tested on the remaining 1/6. Table 2 gives results in both *Exact* and *Indicative* evaluation mode. The baseline values are to be compared with the indicative evaluation, as also partial results were counted. The HyFex results are detailed further into a figure for all mentions of foreign names (526) and one for the subset of name occurrences we could determine to be pronounced according to the German or English languages (180). The latter figures are better – and the improvement over the baseline is much larger – because the current similarity metric relies on German and English pronunciations.

| | Precision | Recall | F ($\beta$=1) |
|---|---|---|---|
| Indicative (All) | 77.6% | 87.6% | 82.3% |
| Exact (All) | 63.8% | 72.1% | 67.7% |
| Indic. (DE-EN) | 81.0% | 90.0% | 85.3% |
| Exact (DE-EN) | 68.5% | 76.1% | 72.1% |
| Baseline (All) | 100% | 51.3% | 67.8% |
| Baseline (DE-EN) | 100% | 43.3% | 60.5% |

Table 2: HyFex for All Foreign Names and German or English Foreign Names vs. Baseline.

| | All | DE-EN |
|---|---|---|
| Person names | 200 | 67 |
| Mentions of these | 526 | 180 |
| Person names mentioned $>$ 4x | 28 | 9 |
| Mentions of these | 268 | 99 |

Table 3: Relation of Person Names and Their Mentions. The most frequent names are *Clinton* (44), *Arafat* (22), and *Yeltsin* (21). The relations for the full set (All) and for the German and English subset (DE-EN) are similar.

Table 3 overviews some relations between person names and their mentions. This property of the corpus is of particular interest since frequent names influence the performance more significantly than rare names. Obviously it is desirable to store frequent names in the gazetteer. Our rule-based system covers some of them, which explains the good

[7] http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp

| NER Systems | Evaluation | No. of PN | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| HyFex Indicative | Foreign PN | 526 | 77.6% | 87.6% | **82.3%** |
| (Chen and Lee, 1996) | | 301 | 76.4% | 76.4% | 76.4% |
| (Gao et al., 2004) | Person Names | unknown | 83.0% | 89.7% | 86.2% |
| (Zhang et al., 2003) | | unknown | 95.5% | 95.7% | 95.6% |

Table 4: Comparison with Results Reported from Other Work

baseline, given the small size of the gazetteer. The statistical component in HyFex can be used to help maintaining the gazetteer over time.

### 6.2. Analysis

Due to the many components integrated in the hybrid system, errors can originate from various sources. The major ones include

- Word segmentation (both systems),
- Language assignment to foreign names
- Conversion to Pinyin,
- Similarity metric defined for SILO.

Investigation on Shanxi word segmentation results showed that among the 526 foreign person names in the People's Daily test dataset, 36 names are segmented incorrectly, and 26 of them could have been correctly recognized and transliterated with the correct segmentation.
It is worth noting that the generation of SAMPA representations for known German and English names with MARY did not bring a significant amount of errors. However, deciding whether a name was German, English, or some other language caused significant problems. We had to rely on the `LANGUAGE` feature each SProUT gazetteer entry is supposed to carry. Unfortunately, many entries did not exhibit such information in a reliable way. Thus, by default, the English phonetic transcription was used, which caused errors for names pronounced according to another language. Errors are either a wrong transliteration result, reducing precision, or a failure to identify the name at all, as the threshold was exceeded, thus reducing recall. On the other hand, it was beneficial in the end to include entries lacking a precise language specification.
To examine the coverage of the Pinyin converter, a Chinese international news text containing 2,197 characters was converted into Pinyin symbols, among which 14 characters could not be recognized, which results in a coverage rate of 99.4%. However, the coverage decreases when the input text contains trigger characters only. Since some of these characters are especially reserved for foreign names and do not belong to the frequently appearing Chinese characters, the converter can only produce a coverage of 95.8%. Of all 526 foreign person names in the test data, there are 10 names (2%) which cannot be completely transcribed.
Moreover, a Chinese character can possibly have several pronunciations depending on the context. This program returns, however, only the most frequently used pronunciation for each character. As far as the experiment shows, this problem does not apparently affect the system performance and is therefore ignored for the time being.
As mentioned in Section 4., a complete list of similarity pairs had to be constructed for SILO. They are defined on the basis of the foreign person names in the corpora used, which amounts to 2,000-2,500 unique names. The similarity metric built up on the basis of this fairly small number of foreign names is very likely to be incomplete, which in turn can mislead SILO and fails to find the best candidate.
In a number of cases, foreign person names were correctly recognized and back-transliterated, but they formed part of a larger unit – e.g., John F. Kennedy airport – and thus were not annotated as foreign person names.
Our system is based on the assumption that standard pronunciation rules be applied to the written names in each language. This may result in different pronunciations for a name in different foreign languages. The comparison with a Chinese pronunciation may produce different similarity results. It is not clear to us whether this contributes to the error rate. If it does, we expect the effect to be small.
Among the various error sources, the sparsity of unique language assignment and the deficiencies of the similarity metric probably have the largest impact on the overall result. Zhang (2007) gives more details on the errors found.

### 6.3. Comparison to other work

A comparison of our results with other work seems difficult due to the different performance of tokenizers, the use of different corpora in different systems, and different evaluation principles. Moreover, most other systems do not separate foreign person names recognition from the person names recognition task in general, and they do not deal with back-transliteration as part of NER.
Table 4 overviews some key results in comparison to some of the work sketched in Section 3. Of course they require some qualification. The system by Chen and Lee (1996) distinguishes the recognition of the foreign person names from the Chinese names. Their test corpus is based on newspapers and seems comparable to ours. As no back-transliteration is carried out, we would compare our indicative results with their best result.
More recent work by Gao et al. (2004) uses a carefully annotated corpus of 40 million Chinese characters, with tests carried out on half a million characters. They do not separate results on foreign person names from those on Chinese person names, which certainly is a reason for showing slightly better results.
The best results reported to our knowledge are from Zhang et al. (2003). They used five months of the People's Daily newspaper for their purely statistic approach. Since they, too, did not give a separate evaluation for the transliterated foreign names, a meaningful comparison of that system with ours remains difficult.
However, it is evident that, different from the above-mentioned work, our system is capable of back-transliterating foreign names and of delivering structured

results that may include additional information related to the person, which can be extracted from the text or a knowledge source (see Figure 4). This flexibility allows for better disambiguation – cf. the *pIrs* example – and renders the system amenable to diverse kind of applications, such as identity tracking.

## 7. Conclusions and Further Work

We described a new, hybrid approach for the recognition of transliterated foreign person names in Chinese that is based on the similarity of phonetic representations of both the foreign and the Chinese version. Dedicated advantages of the implemented system HyFex include the transliteration of foreign person names into their original script and the possibility to generate structured results that may include additional information related to the person.

The F-measure of HyFex measured on publicly available data is up to 85.3% for German and English person names, improving on a gazetteer-based baseline by up to 17.5%.

To improve the performance further, alternative tokenizers and Pinyin converters will be investigated. Improvements of the similarity metric, either manually or by virtue of machine learning techniques, will certainly pay off.

The observation that some trigger characters are more likely to occur in a foreign name than others (Sproat et al., 1996) is not reflected in the system so far. Using statistical information about the distribution of trigger characters should increase the confidence with which a foreign name is located.

The statistical component will be extended to cover multiple results. as described in Section 5.5. It shall also recognize other NE types, such as organization and brand names. Person names will then be recognized with better precision and will support the recognition of some organizations or locations. For instance "John Kennedy" in "John Kennedy International Airport" would then be recognized as part of a location name.

Finally it should be noted that the recognition of foreign names from other languages than German and English is easily possible as soon as phonetic representations are created for them, using suitable TTS tools.

## Acknowledgment

This work has been partially supported by the European Union, Information Society Technologies, under contract no. FP6-027685 to the project MESH. We wish to thank Hans Uszkoreit for pointing out to us a stimulating research task. Many thanks go to members of the DFKI Language Technology lab, in particular Witold Drozdzynski, Andreas Eisele, Ulrich Schäfer and Marc Schröder, for technical support on using the different system components.

## 8. References

Hsin-Hsi Chen and Jen-Chang Lee. 1996. Identification and classification of proper nouns in Chinese texts. *Proc. 16th COLING*, pages 222–229.

Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. Shallow processing with unification and typed feature structures – foundations and applications. *Künstliche Intelligenz*, 1:17–23.

Andreas Eisele and Tim vor der Brück. 2004. Error-tolerant finite-state lookup for trademark search. In *Proc. 27th Annual German Conference on AI*, pages 20–24. Springer.

Jianfeng Gao, Mu Li, Andi Wu, and Changning Huang. 2004. Chinese word segmentation: A pragmatic approach. Technical Report MSR-TR-2004-123, Microsoft.

James L. Hieronyms. 1997. Worldbet phonetic symbols for multilanguage speech recognition and synthesis. Technical report, AT&T Bell Labs.

Chu-Ren Huang, Petr Šimon, and Shu-Kai Hsieh. 2007. Automatic discovery of named entity variants. *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 153 – 156.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Hans-Ulrich Krieger, Witold Drozdzynski, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. A bag of useful techniques for unification-based finite-state transducers. In Ernst Buchberger, editor, *Proceedings of 7th KONVENS*, pages 105–112, 9.

Chun-Jen Lee, Jason S. Chang, and Jyh-Shing Roger Jang. 2006. Extraction of transliteration pairs from parallel corpora using a statistical transliteration model. *Information Sciences*, 176(1):67–90.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007. Semantic transliteration of personal names. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 120–127.

Wei-Hao Lin and Hsin-Hsi Chen. 2002. Backward machine transliteration by learning phonetic similarity. *Proc. 6th Conference on Natural Language Learning*, pages 1–7.

Kaiying Liu. 2001. Research of automatic Chinese word segmentation. In *Int. Workshop on Innovative Language Technology and Chinese Information Processing (ILT&CIP)*, Shanghai, PRC.

Marc Schröder and Jürgen Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365 – 377.

Richard Sproat, Chilin Chih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22:377–404.

John Wells. 1997. SAMPA computer readable phonetic alphabet. In Dafydd Gibbon, Roger Moore, and Richard Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin.

Xinhua News Agency. 1999. *German Name Transliteration Manual*. Commercial Press, Peking, China.

Binyong Yin and Mary Felley. 1990. *Chinese Romanization: Pronunciation and Orthography*. Sinolingua, Bei-

jing.

Huaping Zhang, L. Qun, Hongkui Yu, Xueqi Cheng, and Shuo Bai. 2003. Chinese named entity recognition using role model. *Computational Linguistics and Chinese Language Processing*, 8:29–60.

Yajing Zhang. 2007. Extraction of foreign person names for Mandarin Chinese. Master's thesis, Saarland University.