Identifying Foreign Person Names in Chinese Text

Stephan Busemann, Yajing Zhang
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken

stephan.busemann@dfki.de
yajing.zhang@dfki.de

Motivation

… 路德维希·凡·贝多芬 …

- **Is this a foreign (= non-Chinese) person name (FN)?**
- **What name does it correspond to in Latin script?**

Motivation

… 路德维希·凡·贝多芬 …

- **Is this a foreign (= non-Chinese) person name (FN)?**
- **What name does it correspond to in Latin script?**

Ludwig van Beethoven

Motivation

… 路德维希·凡·贝多芬 …

- **Is this a foreign (= non-Chinese) person name (FN)?**
- **What name does it correspond to in Latin script?**

Ludwig van Beethoven

- **Sample Applications**
  - Machine translation
  - Cross-lingual information extraction
  - Text alignment

Issues of (Back-)Transliteration

Issues of (Back-)Transliteration

- **Transliteration is not a function, e.g. *si***

丝	si1	silk
思	si1	thinking
死	si3	die
饲	si4	feed

Issues of (Back-)Transliteration

丝 **si1 silk**

思 **si1 thinking**

死 **si3 die**

伺 **si4 feed**

- **Transliteration is not a function, e.g. *si***

- **FNs may have multiple encodings, e.g. *Clinton***

柯林顿 **ke1-lin2-dun4 (Taiwan)**

克林顿 **ke4-lin2-dun4 (Mainland)**

Issues of (Back-)Transliteration

丝 **si1 silk**
思 **si1 thinking**
死 **si3 die**
伺 **si4 feed**

- **Transliteration is not a function, e.g. *si***

- **FNs may have multiple encodings, e.g. *Clinton***

柯林顿 **ke1-lin2-dun4 (Taiwan)**
克林顿 **ke4-lin2-dun4 (Mainland)**

- **Final consonants may be omitted, e.g. *Mubarak***

穆巴拉克 **mu4-ba1-la1-ke4**
穆巴拉 **mu4-ba1-la1**

Issues of (Back-)Transliteration

丝 si1 silk
思 si1 thinking
死 si3 die
伺 si4 feed

- **Transliteration is not a function, e.g. *si***

- **FNs may have multiple encodings, e.g. *Clinton***

柯林顿 ke1-lin2-dun4 (Taiwan)
克林顿 ke4-lin2-dun4 (Mainland)

- **Final consonants may be omitted, e.g. *Mubarak***

穆巴拉克 mu4-ba1-la1-ke4
穆巴拉 mu4-ba1-la1

- **Phonetic similarity may be judged differently, e.g. *da Vinci***

达芬奇 da2-fen1-qi2
达文西 da2-wen2-xi1

Issues of (Back-)Transliteration

丝 si1 silk
思 si1 thinking
死 si3 die
伺 si4 feed

- **Transliteration is not a function, e.g. *si***

- **FNs may have multiple encodings, e.g. *Clinton***

柯林顿 ke1-lin2-dun4 (Taiwan)
克林顿 ke4-lin2-dun4 (Mainland)

- **Final consonants may be omitted, e.g. *Mubarak***

穆巴拉克 mu4-ba1-la1-ke4
穆巴拉 mu4-ba1-la1

- **Phonetic similarity may be judged differently, e.g. *da Vinci***

达芬奇 da2-fen1-qi2
达文西 da2-wen2-xi1

- **Pronunciation depends on the origin of the FN, e.g. *Jean***

简 jian3 (EN)
让 rang4 (FR)

Addressing the Task

- **Basic Idea: choose a hybrid approach**
  - Reuse a large gazetteer of FNs in Latin script as a part of a rule-based NER system
  - Integrate a statistical component to automatically back-transliterate FNs into Latin script

Addressing the Task

- **Basic Idea: choose a hybrid approach**
  - Reuse a large gazetteer of FNs in Latin script as a part of a rule-based NER system
  - Integrate a statistical component to automatically back-transliterate FNs into Latin script
- **Coverage**
  - All issues listed, for Simplified Chinese as used in Mainland China
  - Currently FNs pronounced in English and German

Addressing the Task

- **Basic Idea: choose a hybrid approach**
  - Reuse a large gazetteer of FNs in Latin script as a part of a rule-based NER system
  - Integrate a statistical component to automatically back-transliterate FNs into Latin script
- **Coverage**
  - All issues listed, for Simplified Chinese as used in Mainland China
  - Currently FNs pronounced in English and German
- **Exceptions to pronunciation-based transliteration**
  - FNs of Japanese, Korean, Chinese minority languages
  - Conventions for frequently written FNs (e.g. *John* 约翰 yue1-han4)
  - To be covered in a gazetteer of FNs in Chinese script

Gazetteers – More than Word Lists

Gazetteers – More than Word Lists

- **Gazetteer of Chinese entities**

约翰 | GTYPE: zh_person_name | LATIN: “John“

斯 | GTYPE: zh_trigger

经济学家 | GTYPE: zh_position | PROFESSION: "Economist"

Gazetteers – More than Word Lists

- **Gazetteer of Chinese entities**

约翰 | GTYPE: zh_person_name | LATIN: “John“

斯 | GTYPE: zh_trigger

经济学家 | GTYPE: zh_position | PROFESSION: "Economist"

- **Gazetteer of FNs and their pronunciations (SAMPA)**

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

SAMPA created for EN and DE by the TTS system MARY (Schröder and Trouvain, 2001)

Relating a Sequence of Characters to FNs

- **Create Pinyin representation (PR) for a candidate sequence of Chinese characters (CS)**

Relating a Sequence of Characters to FNs

- **Create Pinyin representation (PR) for a candidate sequence of Chinese characters (CS)**

- **Compare PR with all SAMPA phonetic representations (SPRs)**

Relating a Sequence of Characters to FNs

- **Create Pinyin representation (PR) for a candidate sequence of Chinese characters (CS)**
- **Compare PR with all SAMPA phonetic representations (SPRs)**
- **Return**
  - Name string associated with most similar SPR, or
  - State that CS is no FN

„Trigger“ Characters

„Trigger“ Characters

- **Chinese characters used for FNs are limited**

„Trigger“ Characters

- **Chinese characters used for FNs are limited**
- **Sets used in related work were unavailable to us**

„Trigger“ Characters

- **Chinese characters used for FNs are limited**
- **Sets used in related work were unavailable to us**
- **Defined a language-neutral set of characters**
  - Gazetteer of FNs written in Chinese
  - Included additional characters from a person name translation manual (Xinhua News Agency)
  - Removed some ambiguous characters not typical for FNs, sacrificing some recall and gaining much in precision
  - Ended up with a set of 353 characters

„Trigger“ Characters

- **Chinese characters used for FNs are limited**
- **Sets used in related work were unavailable to us**
- **Defined a language-neutral set of characters**
  - Gazetteer of FNs written in Chinese
  - Included additional characters from a person name translation manual (Xinhua News Agency)
  - Removed some ambiguous characters not typical for FNs, sacrificing some recall and gaining much in precision
  - Ended up with a set of 353 characters
- **A FN consists of at least two and at most seven trigger characters**

Comparing Phonetic Similarity with SILO

(Eisele and vor der Brück, 2004)

- **Calculate edit distance based on a metric**
- **Try transducing a Pinyin sign into any of the SAMPA FN representations (FST)**
- **Rank results according to costs**
- **Return the cheapest sign if costs don‘t exceed a threshold**

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

Comparing Phonetic Similarity with SILO

(Eisele and vor der Brück, 2004)

- **Calculate edit distance based on a metric**
- **Try transducing a Pinyin sign into any of the SAMPA FN representations (FST)**
- **Rank results according to costs**
- **Return the cheapest sign if costs don‘t exceed a threshold**

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

Note: Comparing Pinyin with SAMPA rather than with the lexical representation of FNs renders the metric language-neutral.

Back-Transliterating a Candidate Sequence of Chinese Characters into a FN

Chinese	桑普拉斯
Pinyin	**sang1-pu3-la1-si1**
Pinyin	s ang pu l a si
Costs	0.0 0.0 0.2 0.2 0.0 0.0
SAMPA	s {m p r @ s
SAMPA	**s{mpr@s**
Latin	**Sampras**

- **Chinese-to-Pinyin converter (by Jisheng Xie, available from the Internet)**
- **SILO, threshold = 0.4**
- **Gazetteer for FNs and their SAMPA representations**

Back-Transliterating a Candidate Sequence of Chinese Characters into a FN

Chinese	桑普拉斯
Pinyin	**sang1-pu3-la1-si1**
Pinyin	s ang pu l a si
Costs	0.0 0.0 0.2 0.2 0.0 0.0
SAMPA	s {m p r @ s
SAMPA	**s{mpr@s**
Latin	**Sampras**

- **Chinese-to-Pinyin converter (by Jisheng Xie, available from the Internet)**
- **SILO, threshold = 0.4**
- **Gazetteer for FNs and their SAMPA representations**

This describes the statistical component of the hybrid system

The Rule-Based Component: SProUT

- **Shallow parsing system based on typed feature structures**
- **Combines**
  - Morphological analysis,
  - Token information, and
  - Gazetteer information
  - … into rules

The Rule-Based Component: SProUT

- **Shallow parsing system based on typed feature structures**
- **Combines**
  - Morphological analysis,
  - Token information, and
  - Gazetteer information
  - … into rules

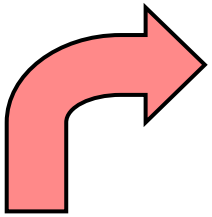
```
foreign_person :>
   gazetteer & [ GTYPE zh_person_position, PROFESSION #position ]?
   gazetteer & [ GTYPE zh_person_name, SURFACE #zh1, LATIN #n1 ]
   gazetteer & [ GTYPE zh_name_separator, SURFACE #sep ]
   gazetteer & [ GTYPE zh_person_name, SURFACE #zh2, LATIN #n2 ]
-> ne-person & [SURFACE #surface,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

Source: Stephan Busemann, Yajing Zhang

(Drozdzynski et al. 2004)

The Rule-Based Component: SProUT

- **Shallow parsing system based on typed feature structures**
- **Combines**
  - Morphological analysis,
  - Token information, and
  - Gazetteer information
  - … into rules



```
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"
```

```
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).
```

Source: Stephan Busemann, Yajing Zhang

(Drozdzynski et al. 2004)

Integration of the Statistical into the Rule-Based Component

Integration of the Statistical into the Rule-Based Component

- **First the gazetteer of Chinese FNs is checked**

Integration of the Statistical into the Rule-Based Component

- First the gazetteer of Chinese FNs is checked
- If it fails, newly designed SProUT rules call a functional operator *CombineStatistics* on a sequence of 2-7 trigger characters

Integration of the Statistical into the Rule-Based Component

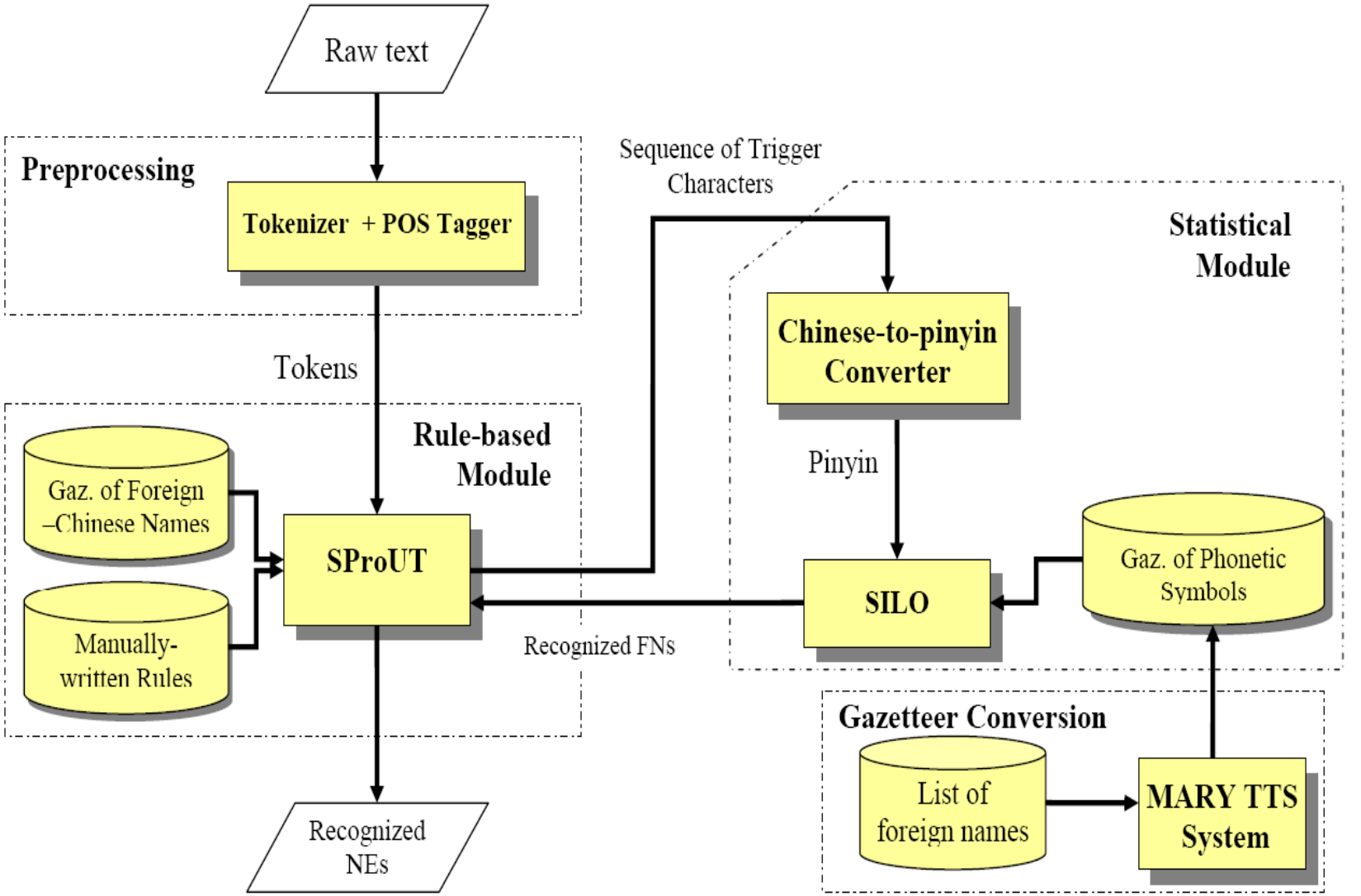
- First the gazetteer of Chinese FNs is checked
- If it fails, newly designed SProUT rules call a functional operator *CombineStatistics* on a sequence of 2-7 trigger characters
- *CombineStatistics* returns a typed feature structure *ne-person* containing a name in Latin script, or it fails

Integration of the Statistical into the Rule-Based Component

- First the gazetteer of Chinese FNs is checked
- If it fails, newly designed SProUT rules call a functional operator *CombineStatistics* on a sequence of 2-7 trigger characters
- *CombineStatistics* returns a typed feature structure *ne-person* containing a name in Latin script, or it fails
- Sample SProUT rule yielding either a first name or a surname

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

The HyFex NER System



Source: Stephan Busemann, Yajing Zhang

Evaluation: Data and Principles

Evaluation: Data and Principles

- **Data**
  - January 1998 issues of *People's Daily* newspaper (publicly available on the Internet with segment annotation)
  - 1.1 million words, FNs predominantly from politics and sports
  - Annotated FNs (180 mentions of 67 EN or DE names)
  - Used 5/6 to tune the HyFex system and 1/6 for test

Evaluation: Data and Principles

- **Data**
  - January 1998 issues of *People's Daily* newspaper (publicly available on the Internet with segment annotation)
  - 1.1 million words, FNs predominantly from politics and sports
  - Annotated FNs (180 mentions of 67 EN or DE names)
  - Used 5/6 to tune the HyFex system and 1/6 for test
- **Principles**
  - **Exact**: found correct sequence and returned correct backtransliteration
  - **Indicative**: FN seen, backtransliteration incorrect, or name only partially recognized

Evaluation: Data and Principles

- **Data**
  - January 1998 issues of *People's Daily* newspaper (publicly available on the Internet with segment annotation)
  - 1.1 million words, FNs predominantly from politics and sports
  - Annotated FNs (180 mentions of 67 EN or DE names)
  - Used 5/6 to tune the HyFex system and 1/6 for test
- **Principles**
  - **Exact**: found correct sequence and returned correct backtransliteration
  - **Indicative**: FN seen, backtransliteration incorrect, or name only partially recognized
- **Baseline: Chinese gazetteer version of SProUT**
  - Records just about 800 frequently used names

Intrinsic Evaluation: Results and Analysis

	Precision	Recall	F (β = 1)
Indicative	81.0	90.0	85.3
Exact	**68.5**	**76.1**	**72.1**
Baseline	100.0	43.3	60.5

Source: Stephan Busemann, Yajing Zhang

Intrinsic Evaluation: Results and Analysis

	Precision	Recall	F (β = 1)
Indicative	81.0	90.0	85.3
Exact	**68.5**	**76.1**	**72.1**
Baseline	100.0	43.3	60.5

- **Major sources of errors**
  - Missing or false language assignment to gazetteer entries
  - Deficiencies in the similarity metric (data sparsity)

Intrinsic Evaluation: Results and Analysis

	Precision	Recall	F (β = 1)
Indicative	81.0	90.0	85.3
Exact	**68.5**	**76.1**	**72.1**
Baseline	100.0	43.3	60.5

- **Major sources of errors**
  - Missing or false language assignment to gazetteer entries
  - Deficiencies in the similarity metric (data sparsity)
- **Other notable sources of errors**
  - Conversion to Pinyin
  - Names in context („John F. Kennedy airport“)

Intrinsic Evaluation: Results and Analysis

	Precision	Recall	F (β = 1)
Indicative	81.0	90.0	85.3
Exact	**68.5**	**76.1**	**72.1**
Baseline	100.0	43.3	60.5

Note: The paper has figures for EN/DE FNs (here) and for all FNs.

- **Major sources of errors**
  - Missing or false language assignment to gazetteer entries
  - Deficiencies in the similarity metric (data sparsity)
- **Other notable sources of errors**
  - Conversion to Pinyin
  - Names in context („John F. Kennedy airport“)

Extrinsic Analysis: Comparison to Some Other Work

Difficult due to different tokenizers, corpora, and system aims.
No information on #mentions / #names.

NER System	Prec.	Recall	F (β = 1)	Remarks
HyFex (Indicative)	77.6	87.6	**82.3**	Fig. for *all* FNs in the gazetteer
Chen/Lee 1996	76.4	76.4	76.4	Corpus also newspaper text. No back-transliteration
Gao et al. 2004	93.0	89.7	86.2	Includes Chinese names.
Zhang et al. 2003	95.5	95.7	95.6	NER ↔ word segmentation. People's Daily. Includes Chinese names. No back-transl.

Source: Stephan Busemann, Yajing Zhang

Conclusions and Further Work

- **Recognition of FNs in Chinese text and back-transliteration according to an extendable resource in Latin script**

Conclusions and Further Work

- **Recognition of FNs in Chinese text and back-transliteration according to an extendable resource in Latin script**
- **Types allow us to search for complex structured information rather than just NEs**

Conclusions and Further Work

- **Recognition of FNs in Chinese text and back-transliteration according to an extendable resource in Latin script**
- **Types allow us to search for complex structured information rather than just NEs**
- **Language-neutral**
- **Adding pronunciations according to another language requires TTS functionality to create SAMPA representations**

Conclusions and Further Work

- **Recognition of FNs in Chinese text and back-transliteration according to an extendable resource in Latin script**
- **Types allow us to search for complex structured information rather than just NEs**
- **Language-neutral**
- **Adding pronunciations according to another language requires TTS functionality to create SAMPA representations**
- **Some possible improvements and extensions**
  - Experiment with other word segmenters and Pinyin converters
  - Tune the SILO metric, preferrably by machine-learning
  - Allow for n best outputs of *CombineStatistics*

Follow-up: Disambiguation by Context

- **Same pronunciation**
  - David Peirce
  - David Pierce
  - David Pearce
- **"Famous economist David Pearce stated that …"**
- **To be implemented - currently *CombineStatistics* only returns just one result**

```
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"
```

Source: Stephan Busemann, Yajing Zhang

Thank You for Your Attention!

HyFex is More than the Sum of Its Parts …

- **Reused software and resources**
  - ShanXi University tokenizer
  - SProUT with gazetteer of Chinese entities (800 FNs)
  - Gazetteer of FNs (85.000 entries)
  - Chinese-to-Pinyin converter
  - SILO
  - MARY TTS system
- **Newly developed software and resources**
  - Set of trigger characters
  - SILO metric for Pinyin to SAMPA
  - Workflow implementation (CombineStatistics)
  - FN corpus annotation

Overview of the Remainder of the Talk

- **Relating Chinese Characters to FNs**
  - Gazetteers
  - Comparing Pinyin and SAMPA
- **Implementation: The HyFex NER System**
- **Evaluation**
- **Conclusions**