

美

**Mandarin-English Information (MEI):
Investigating Translingual Speech Retrieval
Johns Hopkins University Summer Workshop 2000**

*Presented at the
ANLP-NAACL 2000*

Embedded Machine Translation Systems Workshop

The MEI Team

MEI Team

- Senior Members

Helen Meng	Chinese University of Hong Kong
Erika Grams	Advanced Analytic Tools
Sanjeev Khudanpur	Johns Hopkins University
Gina-Anne Levow	University of Maryland
Douglas Oard	University of Maryland
Patrick Schone	US Department of Defense
Hsin-Min Wang	Academia Sinica, Taiwan

- Students

Berlin Chen	National Taiwan University
Wai-Kit Lo	Chinese University of Hong Kong
Karen Tang	Princeton University
Jianqiang Wang	University of Maryland

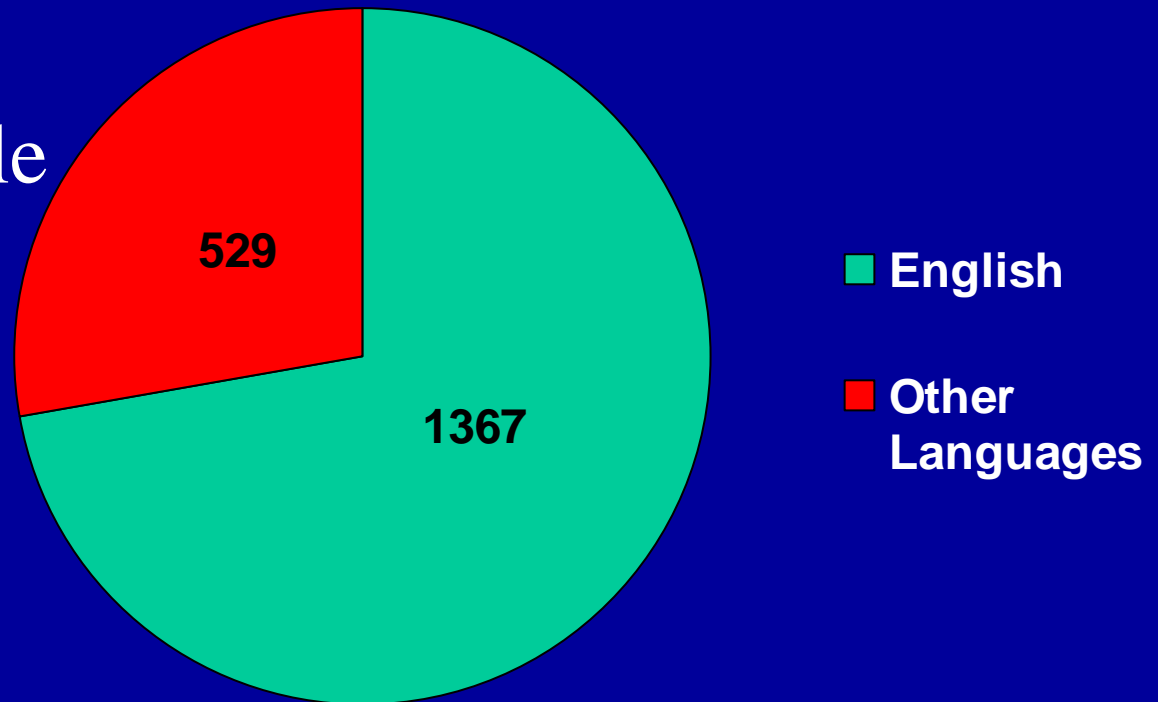
Outline

- Audio indexing
- MEI Project overview
- Research challenges
- System architecture
- Collaboration opportunities

Motivation

- Speech retrieval applications are emerging
 - e.g., <http://speechbot.research.compaq.com>

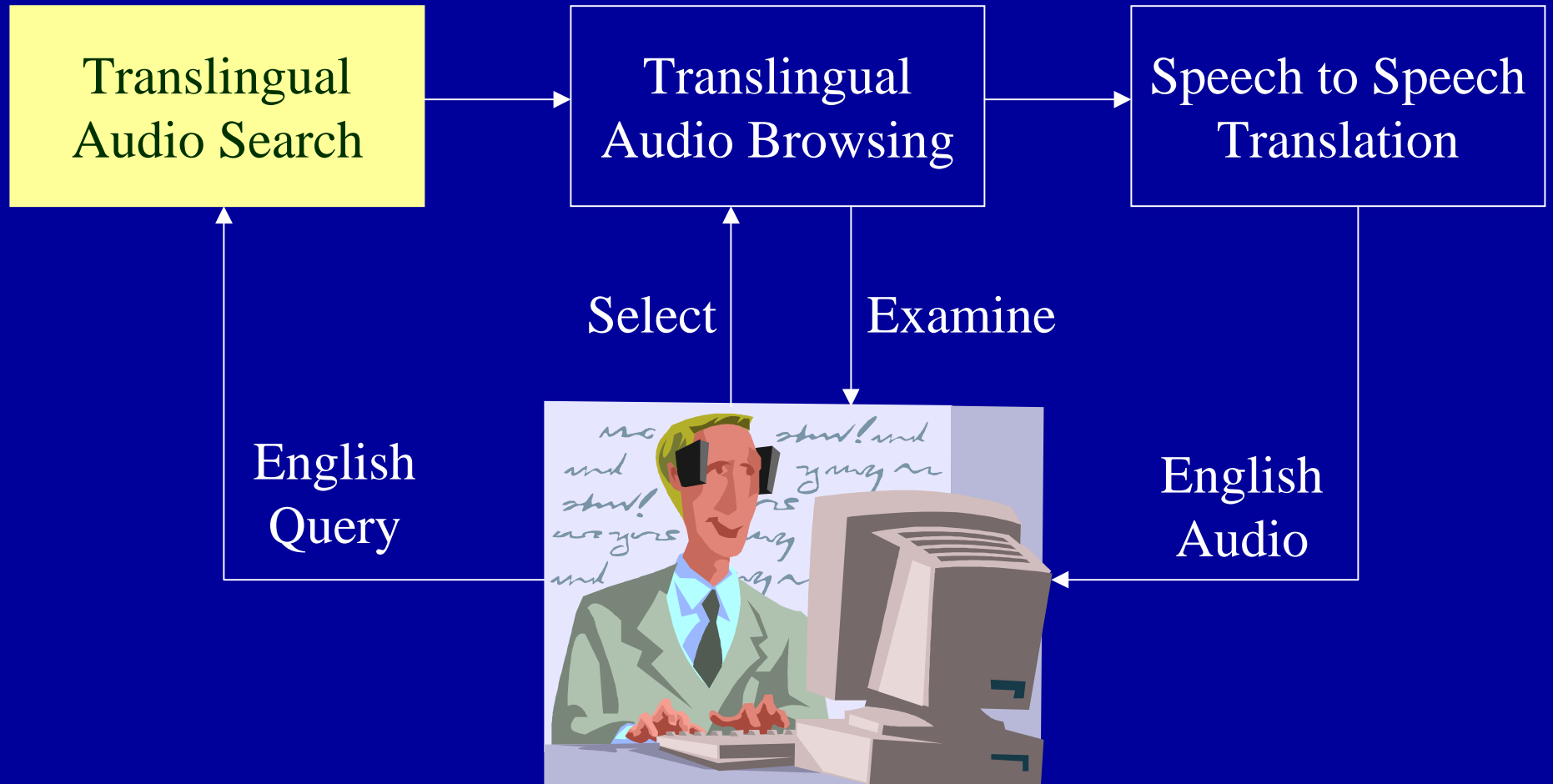
- Internet-accessible
Radio and
Television
Stations



source: www.real.com, Feb 2000

The Big Picture

MEI 美

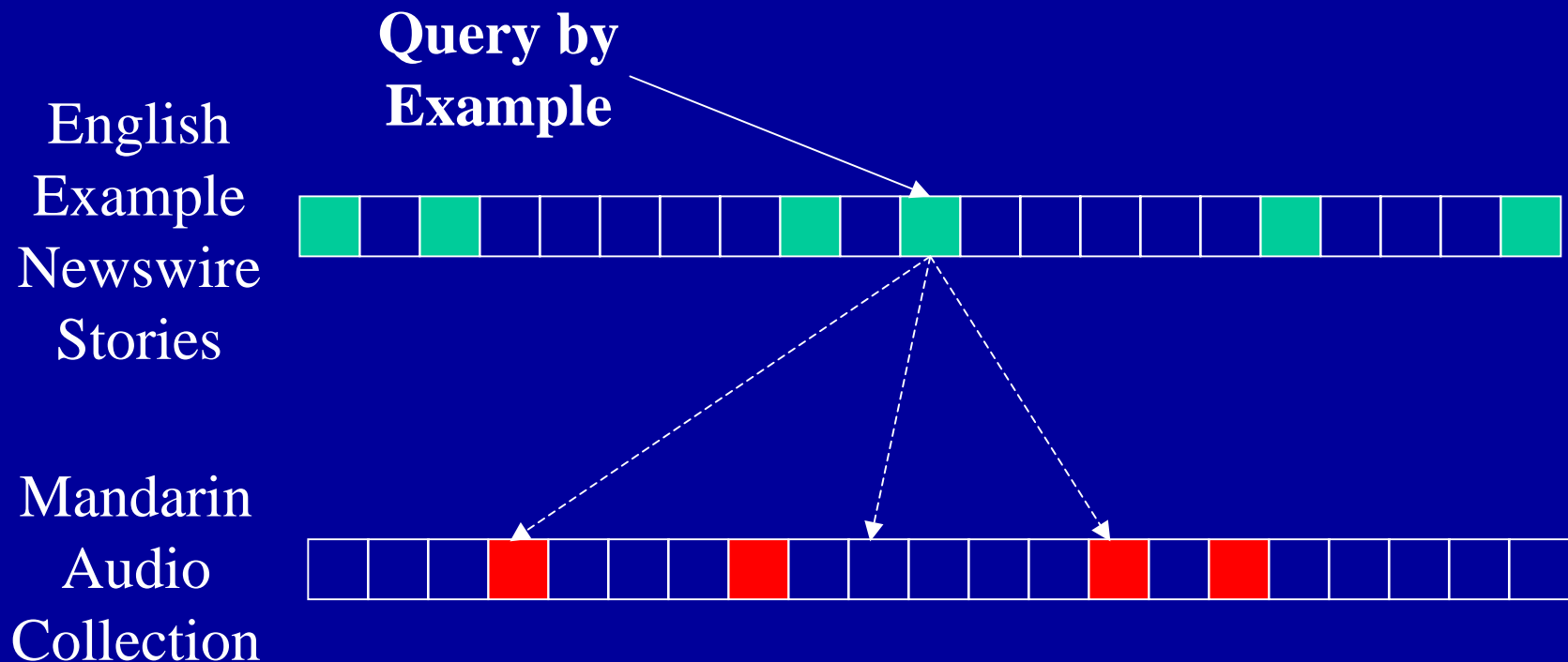


Related Work

- TREC Spoken Document Retrieval
 - Close coupling of recognition and retrieval
- TREC Cross-Language Retrieval
 - Close coupling of translation and retrieval
- TDT-3 Topic Tracking
 - Coupling recognition, translation and retrieval
 - Using speech recognition transcripts

The MEI Project

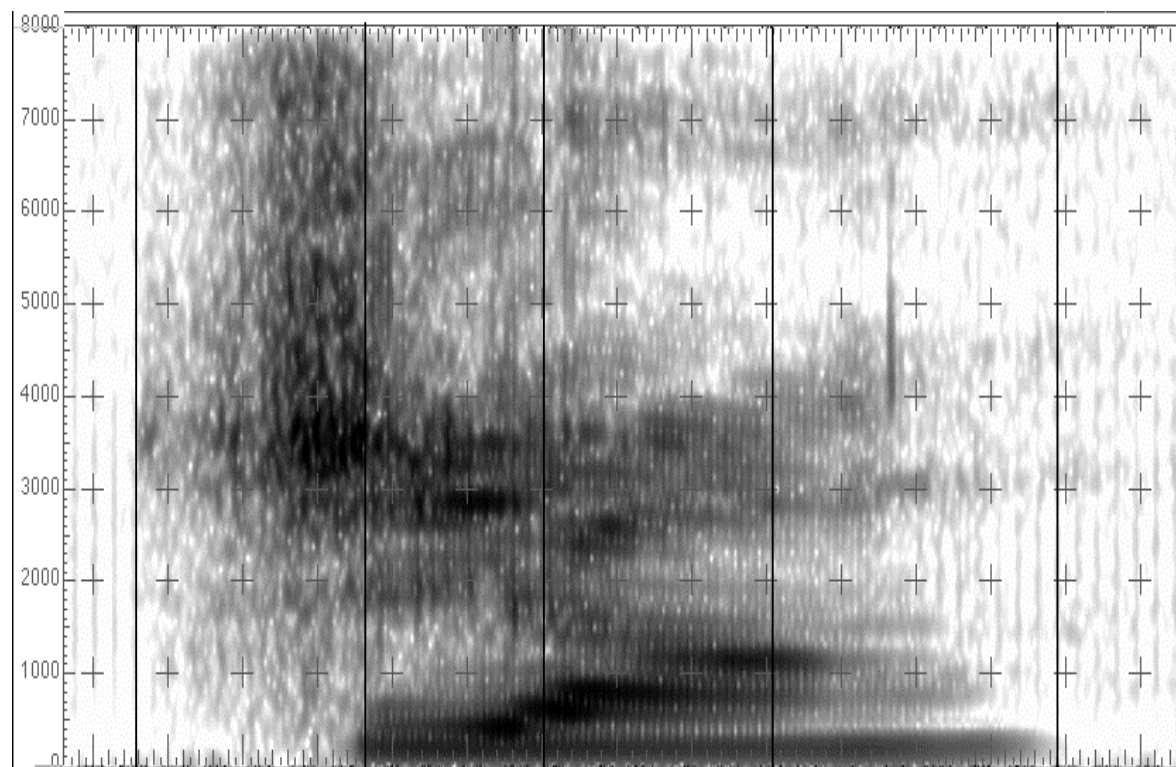
- Closely couple recognition and translation
 - For the purpose of retrieval
- Using English examples, find Mandarin audio



Research Challenges

- Multi-scale audio indexing
 - Multiple feature sets capture more information
- Multi-scale translation
 - Lexicon and pronunciation are complementary
- Multi-scale retrieval
 - Combination of evidence can add robustness

Multi-scale Mandarin Audio Indexing



Preme/Toneme	/j/	/i/	/a/	/ng/
Preme/Core Final	/ji/		/ang/	
Initial/Final	/j/	/iang/		

Multi-scale Translation

- Word-scale
 - Dictionary-based [Levow & Oard 00]
 - Parallel corpora [Nie 99]
 - Comparable corpora [Fung 98]
- Subword-scale [Knight & Graehl 97]
 - Cross-language phonetic mapping
 - 北愛爾蘭 /bei2 ai4 er3 lan2/
 - Kosovo (/ke1-sou3-wo4/, /ke1-sou3-fo2/, /ke1-sou3-fu1/, /ke1-sou3-fu2/)

Cross-Language Phonetic Mapping

- Syllabify English spelling
 - e.g. Jiang Zemin, Shandong Province
- Map English pronunciation to Mandarin
 - Convert phonemes to pinyin
 - e.g. /k ow s ax v ow/ to /ke1-suo3-wo4/
 - Plan to investigate alternative techniques
 - Rule-based
 - Statistical mapping

Multi-scale Retrieval

- Word-scale exploits lexical knowledge
 - Enhances precision
- Subwords can achieve complete coverage
 - Enhances recall
- Combination of evidence may be best
 - If a good merging strategy can be found

Multi-scale Retrieval Techniques

- Subword-scale
 - Syllable lattice matching [Chen, Wang & Lee 00]
 - Overlapping syllable n -grams [Meng et al. 99]
 - Syllable confusion matrix [Meng et al. 99]
- Word-scale
 - Structured queries [Pirkola 98]
 - Structured translation [Sperer & Oard 00]

Merging Strategies

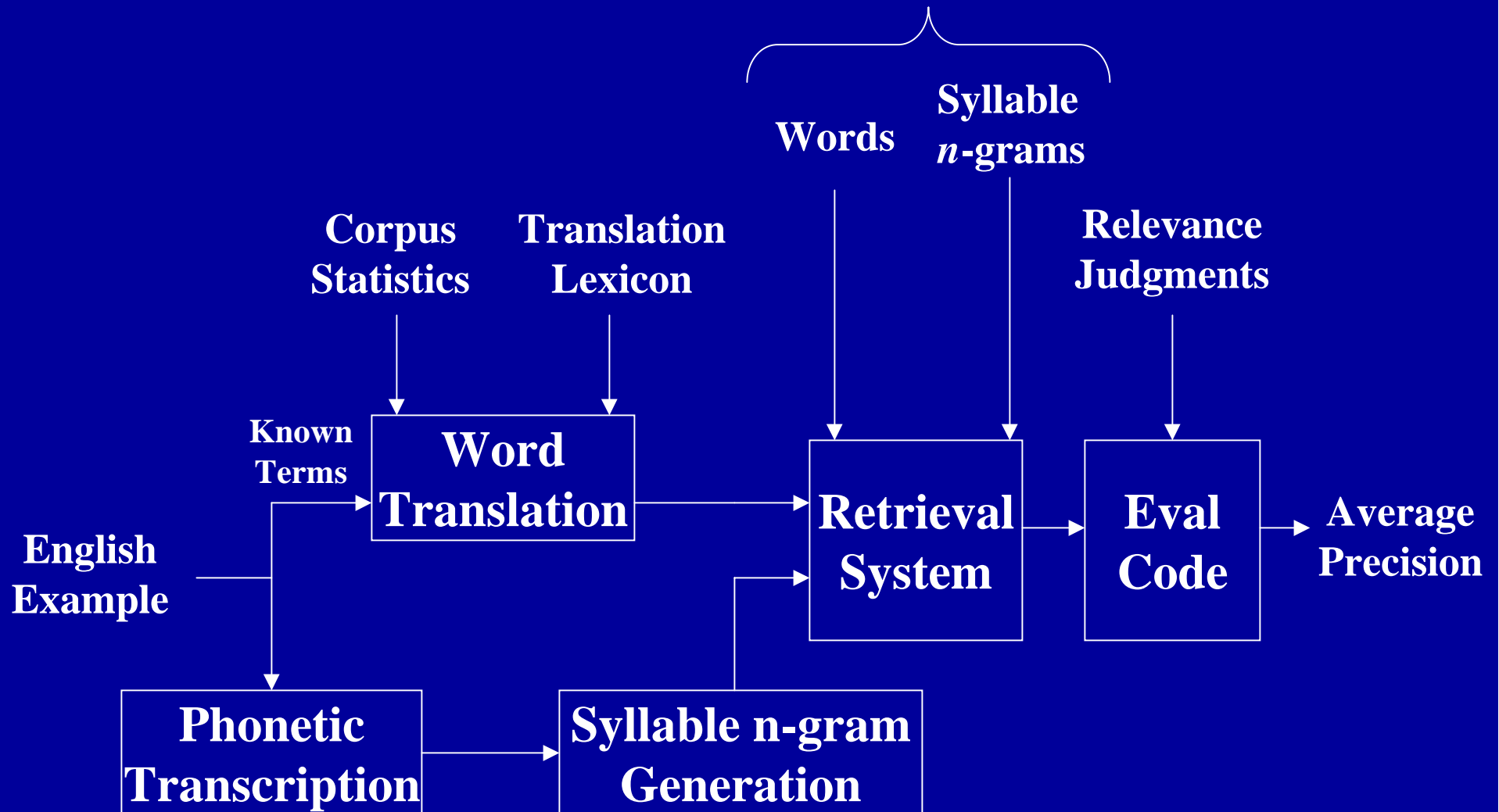
- Loose coupling
 - Separate retrieval runs
 - Merge ranked lists [Voorhees 95]
- Tight coupling [Ng 00]
 - Unified indexing of words and subwords
 - Single ranked list

Robust Retrieval

- Multiple causes
 - Speech recognition errors
 - Translation ambiguity
 - Transliteration ambiguity
- Possible solutions
 - Weighted *n*-best indexing [Levow & Oard 00]
 - Syllable lattice indexing [Chen, Wang & Lee 00]
 - Syllable confusion expansion [Meng et al. 99]
 - Structured queries [Pirkola 98]
 - Document expansion [Levow & Oard 00]

System Architecture Overview

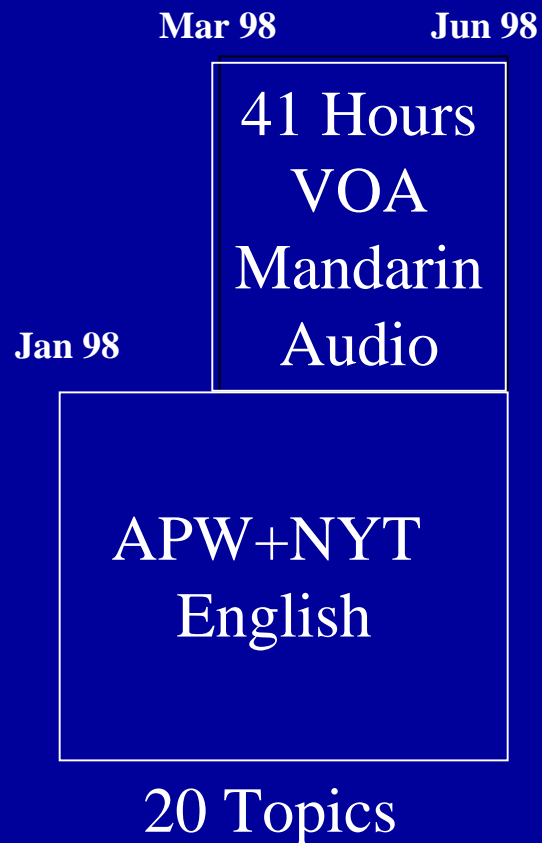
Mandarin Documents



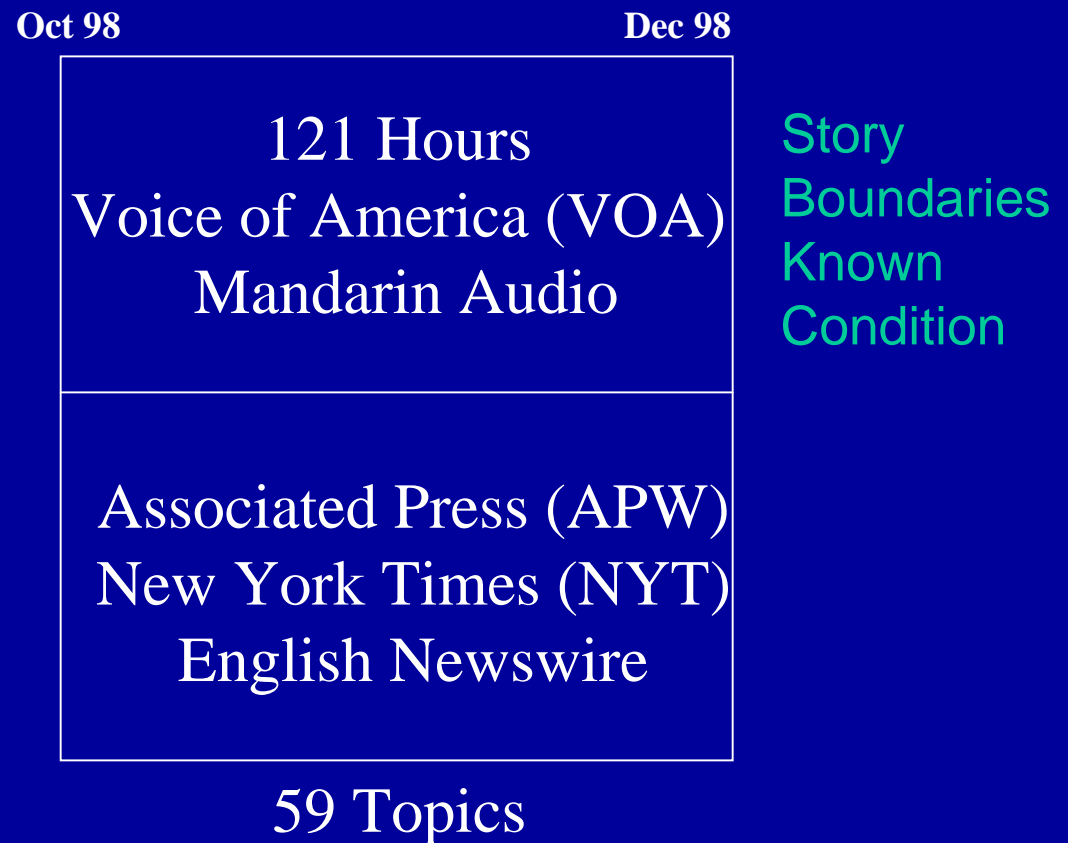
The TDT Collections

- Four stories per topic in each language
 - Each reporting on some aspect of one event

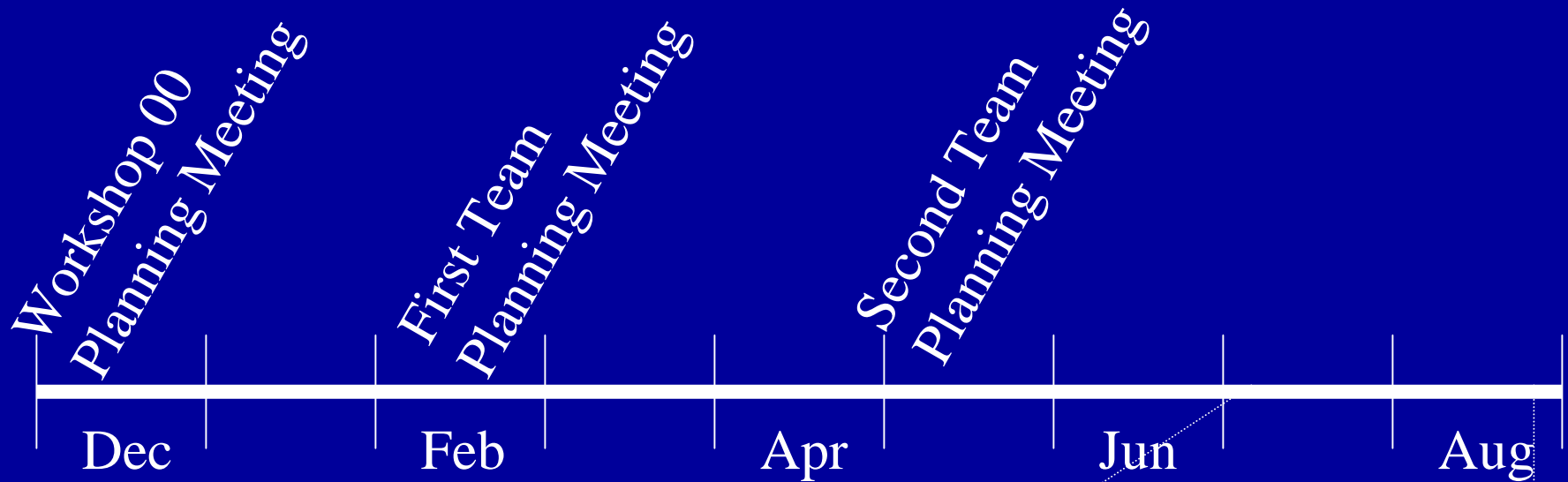
Development Test (TDT-2)



Evaluation (TDT-3)



MEI Project Schedule



Six Weeks at Hopkins:



Things We Need

- Ideas
 - To sharpen our focus
- Connections
 - To build a community of interest
- Resources
 - To build on what others have done

For More Information

- MEI Project
 - <http://www.glue.umd.edu/~meiweb>
- Translingual Retrieval
 - <http://www.clis.umd.edu/dlrg/clir>
- Speech Retrieval
 - <http://www.clis.umd.edu/dlrg/speech>
- Hopkins Summer Workshop Series
 - <http://www.clsp.jhu.edu/workshops>

Backup Slides

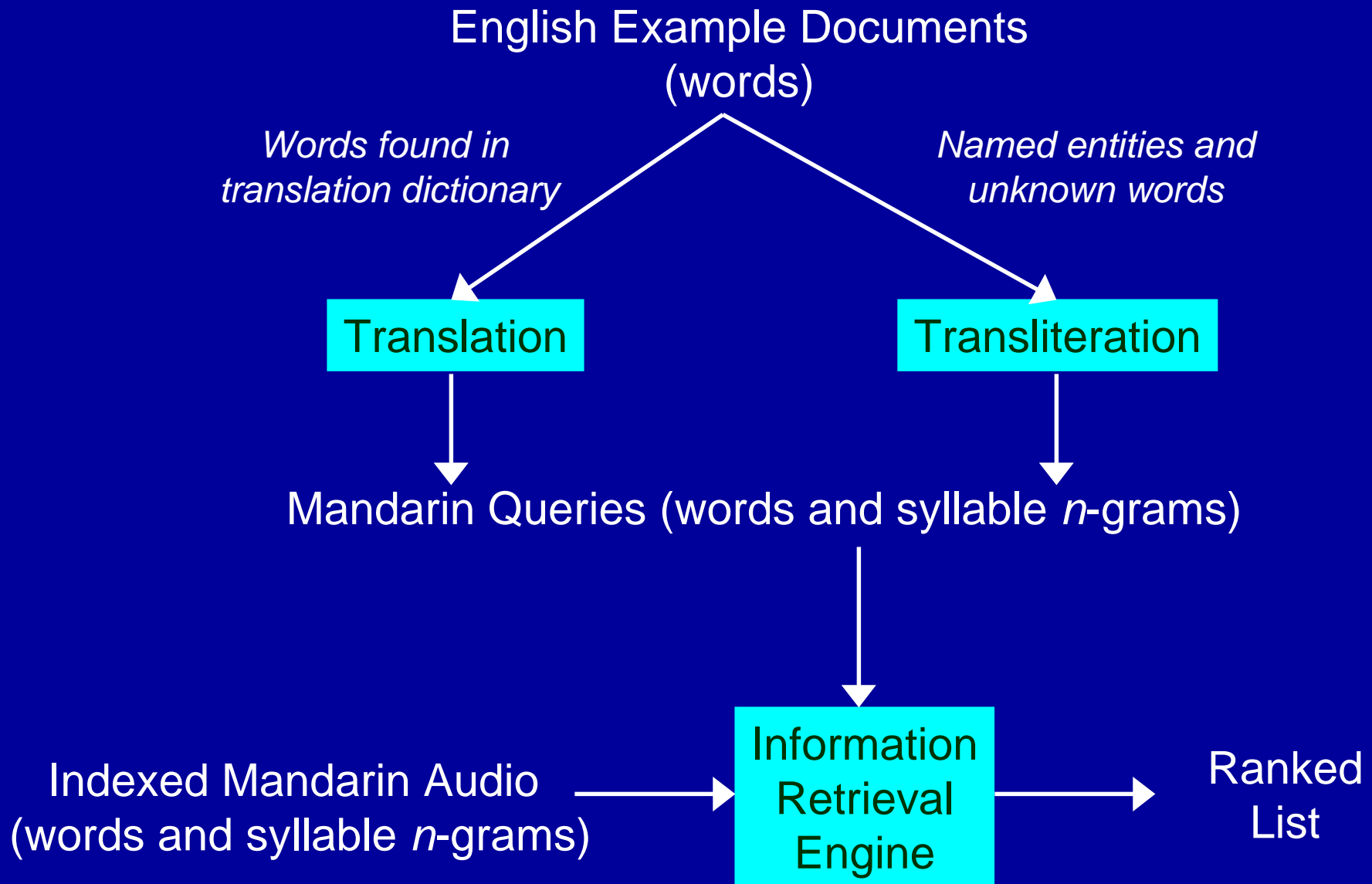
Background: Chinese

- Many dialects (e.g., Mandarin and Cantonese)
 - differences in phonetics, vocabularies, syntax...
- Syllable-based language
 - ~400 base syllables, 4 lexical tones + light tone
- Syllable structure **(CG)V(X)**
 - **(CG)**: onset, optional, consonant+medial glide
 - **V**: nuclear vowel
 - **X**: coda, glide / alveolar nasal / velar nasal
 - ~ 21 initials, 39 finals

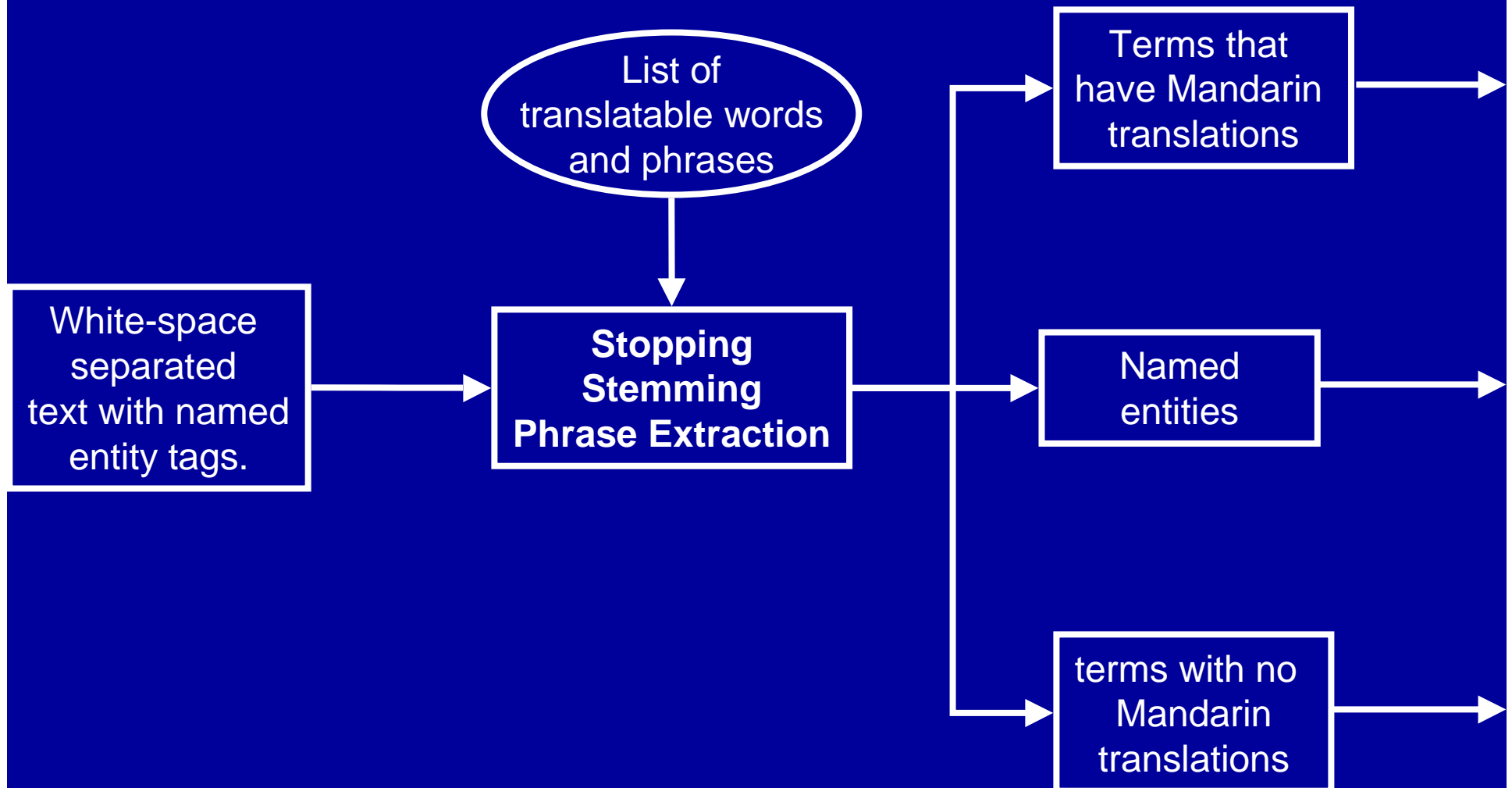
Background: Chinese (cont)

- Characters (written) -> syllables (spoken)
 - Degenerate mapping
 - 行 /hang2/, /hang4/, /heng2/ or /xing2/
 - /fu4 shu4/ (LDC's CALLHOME lexicon)
- 富庶 負數 復數 覆述
- Tokenization / Segmentation
 - /zhe4 yi1 wan3 hui4 ru2 chang2 ju3 xing2/

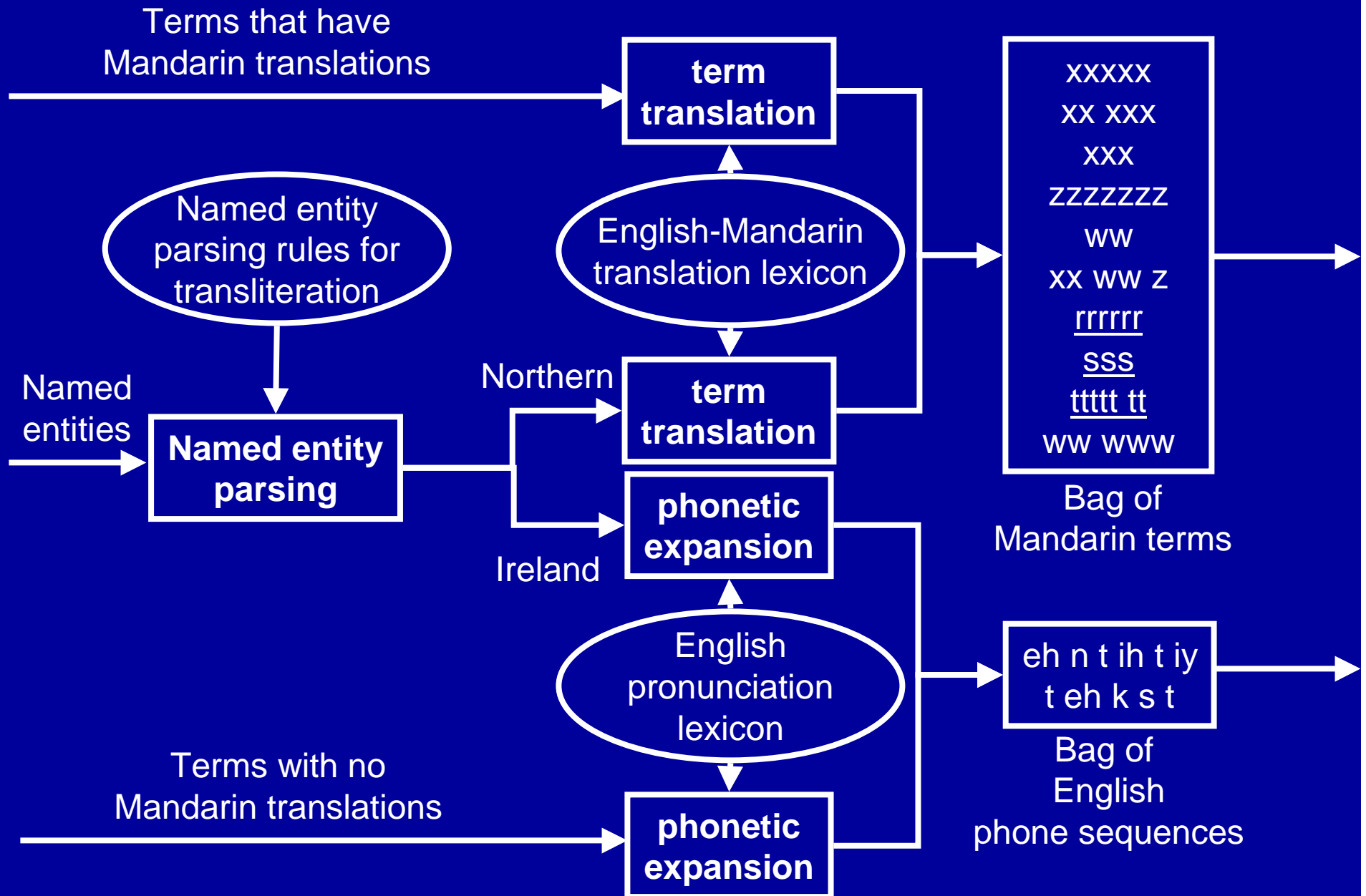
Translating the English Example



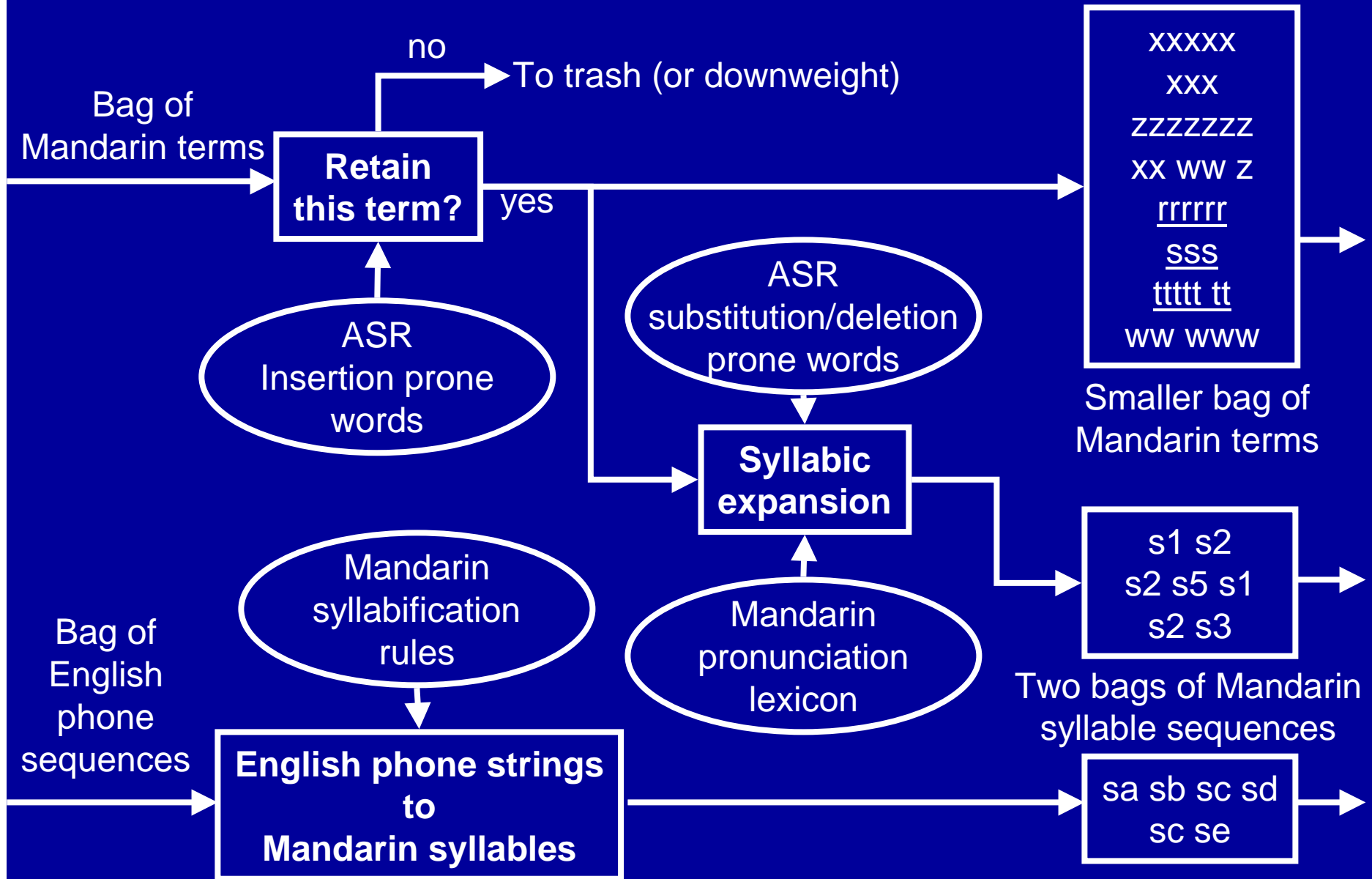
Detailed Query Processing (1)



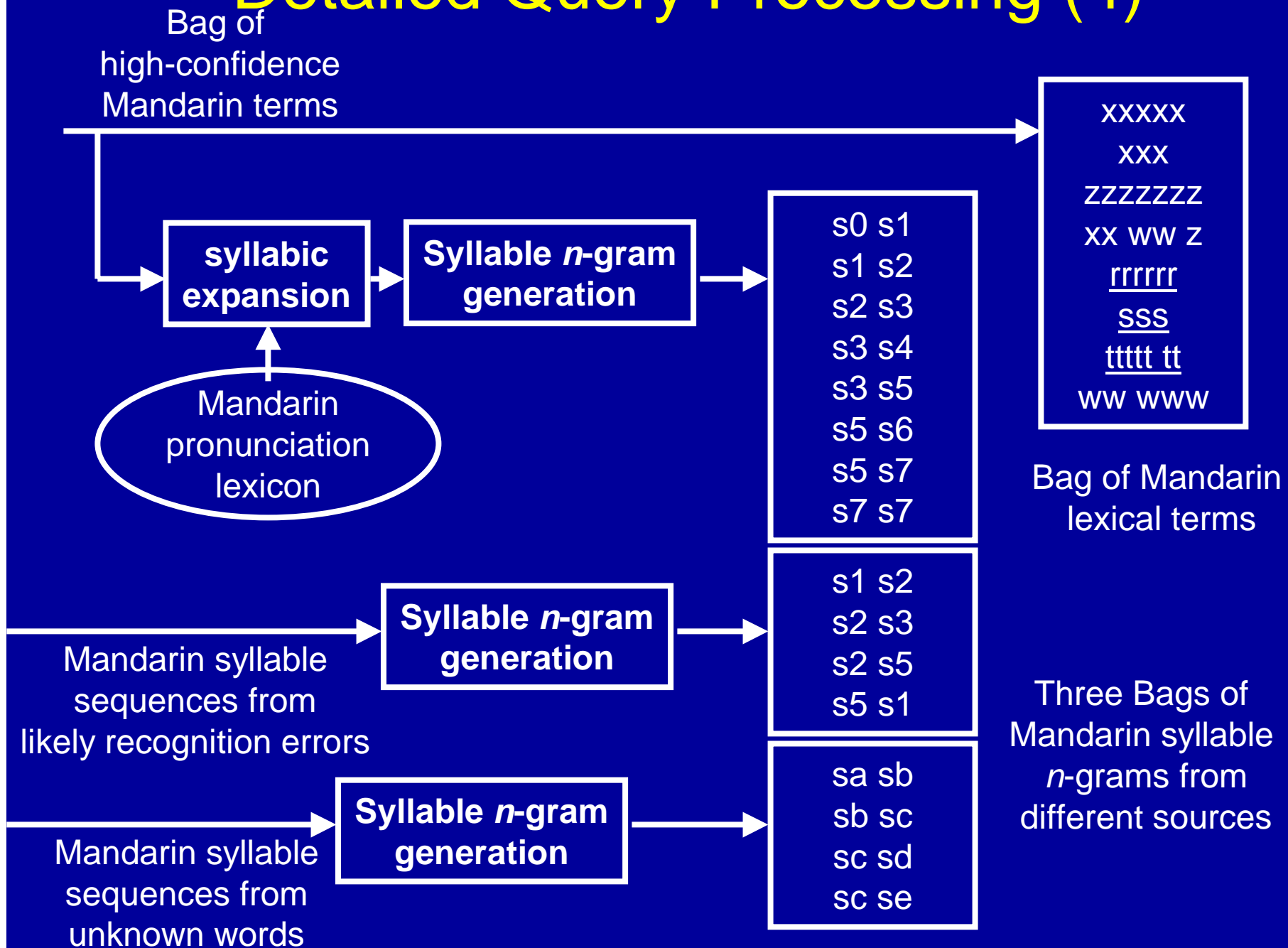
Detailed Query Processing (2)



Detailed Query Processing (3)



Detailed Query Processing (4)



Translating the Mandarin Audio

