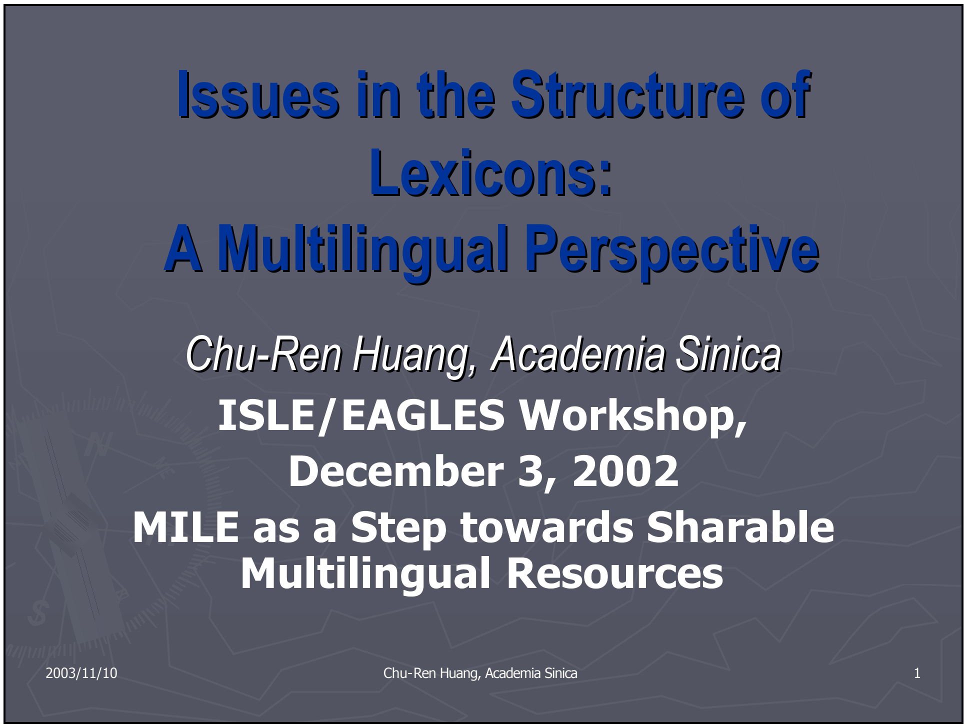
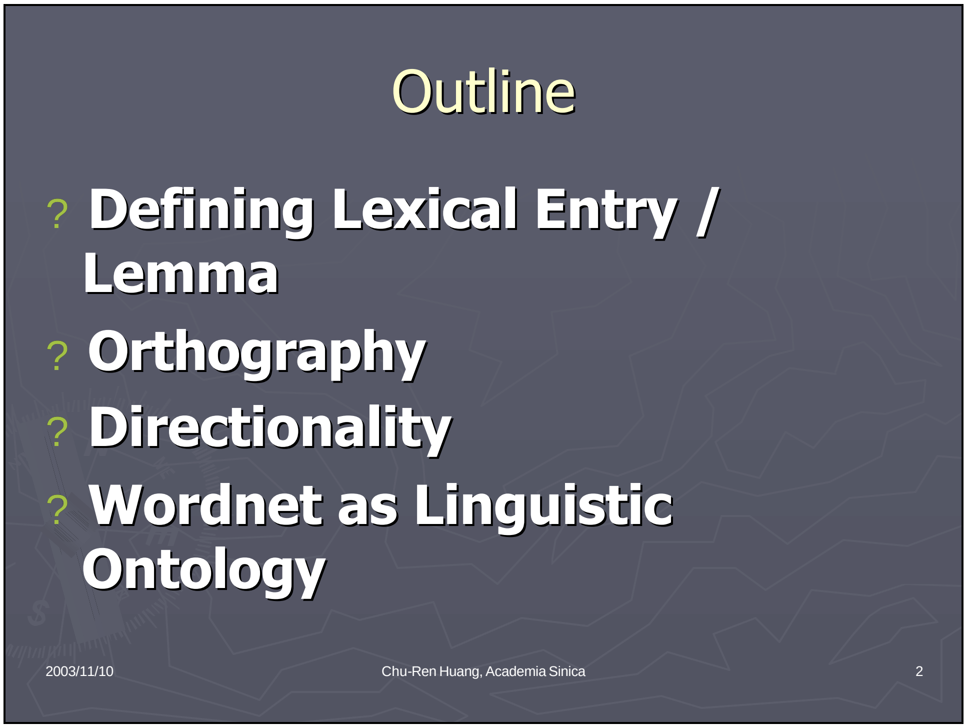


Issues in the Structure of Lexicons: A Multilingual Perspective

Chu-Ren Huang, Academia Sinica

ISLE/EAGLES Workshop, December 3, 2002

MILE as a Step towards Sharable Multilingual Resources



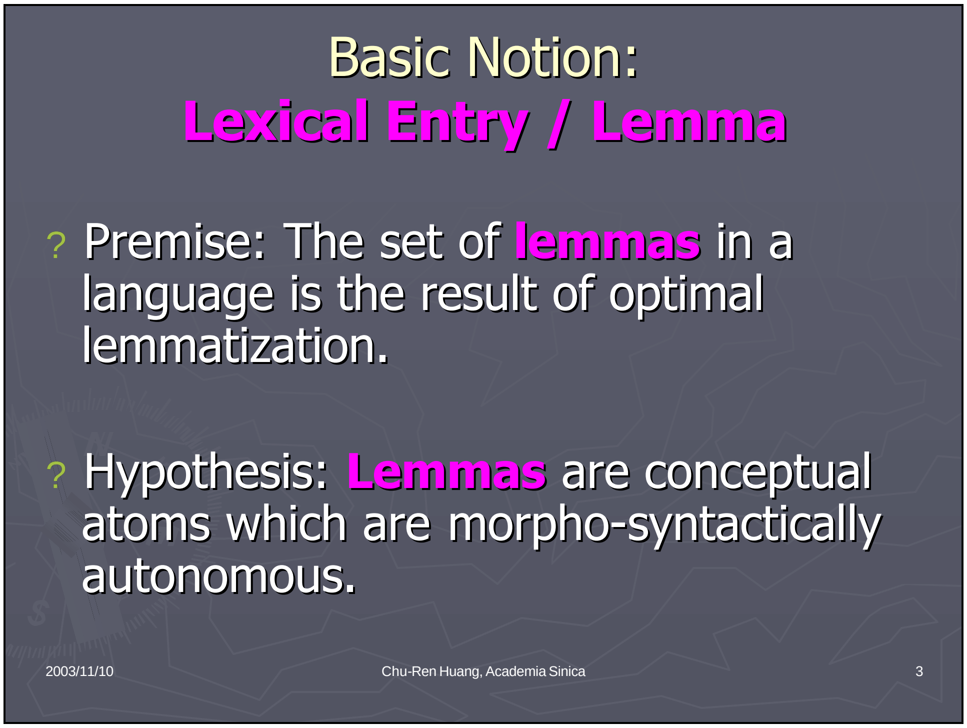
Outline

? **Defining Lexical Entry / Lemma**

? **Orthography**

? **Directionality**

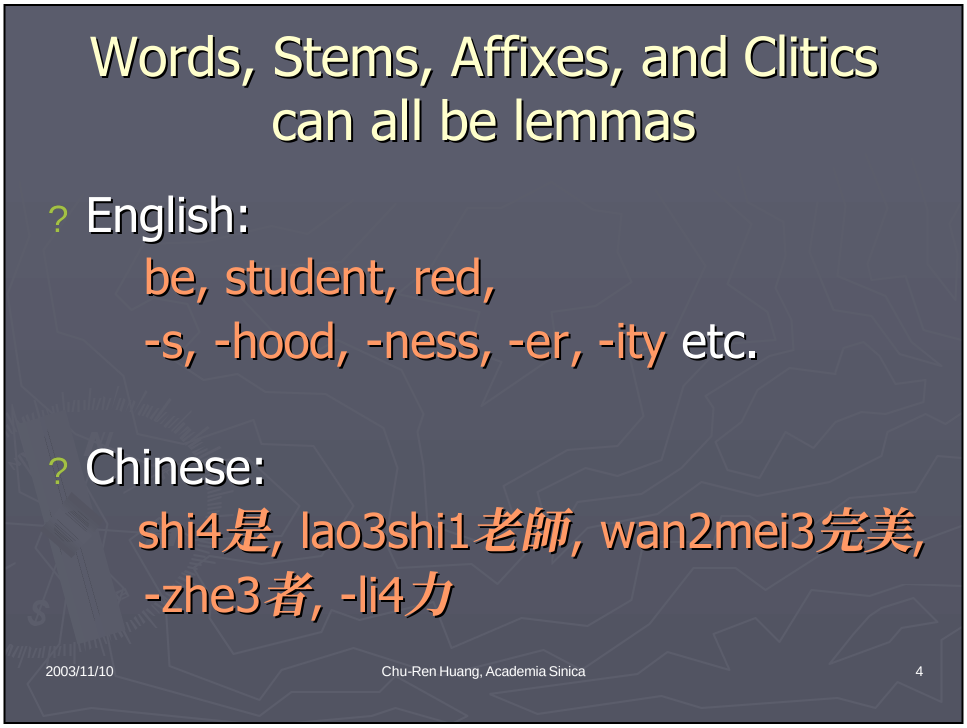
? **Wordnet as Linguistic Ontology**



Basic Notion: **Lexical Entry / Lemma**

? Premise: The set of **lemmas** in a language is the result of optimal lemmatization.

? Hypothesis: **Lemmas** are conceptual atoms which are morpho-syntactically autonomous.



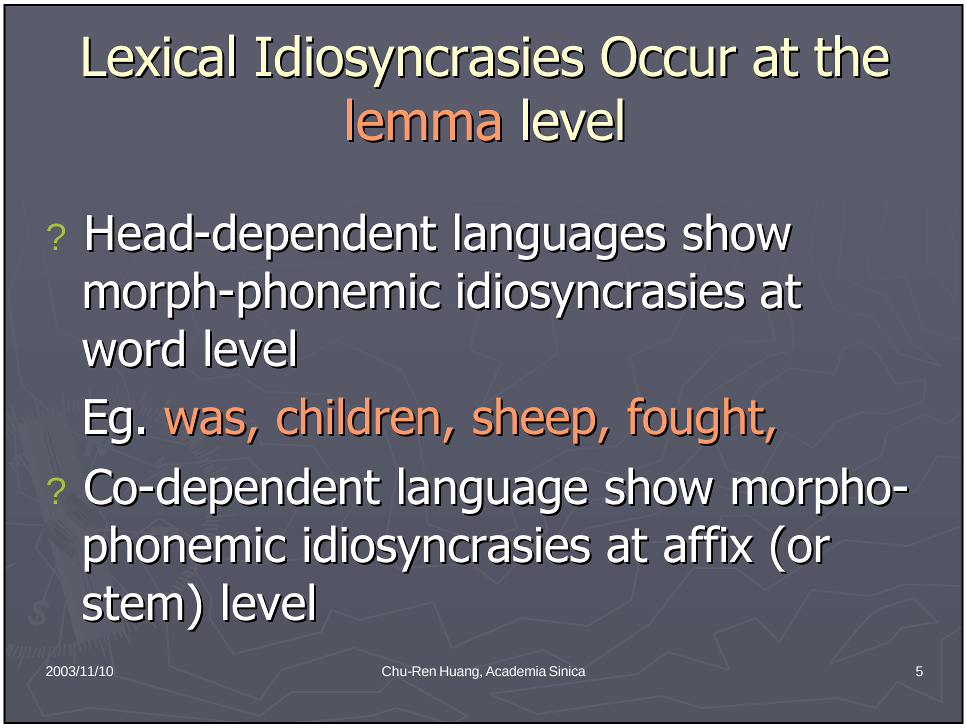
Words, Stems, Affixes, and Clitics can all be lemmas

? English:

be, student, red,
-s, -hood, -ness, -er, -ity etc.

? Chinese:

shi4是, lao3shi1老師, wan2mei3完美,
-zhe3者, -li4力



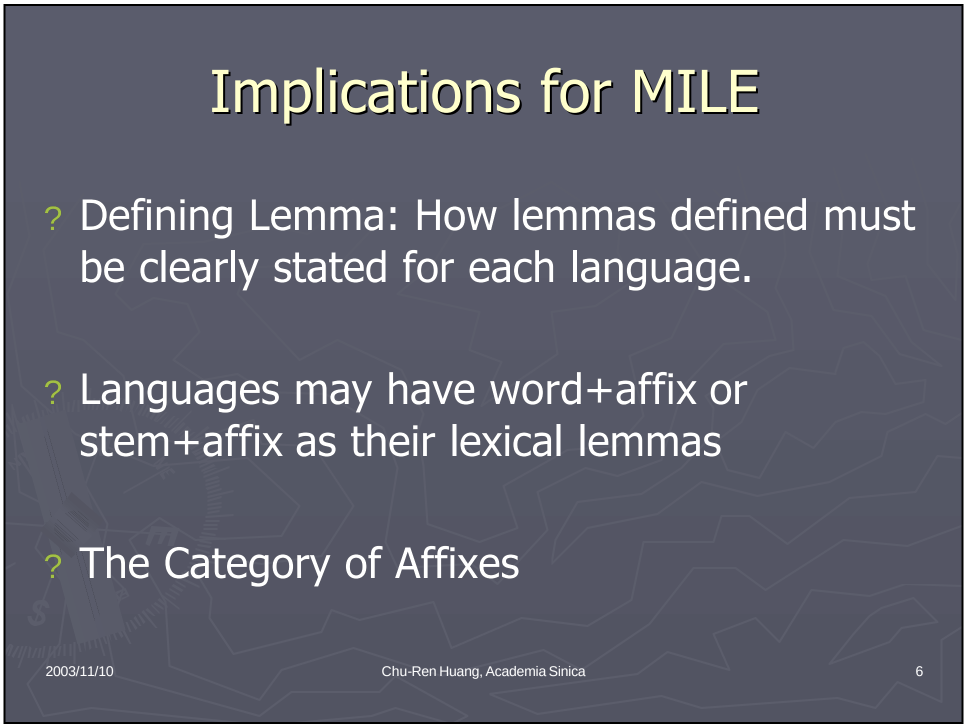
Lexical Idiosyncrasies Occur at the lemma level

? Head-dependent languages show morph-phonemic idiosyncrasies at word level

Eg. was, children, sheep, fought,

? Co-dependent language show morpho-phonemic idiosyncrasies at affix (or stem) level

2003/11/10 Chu-Ren Huang, Academia Sinica 5



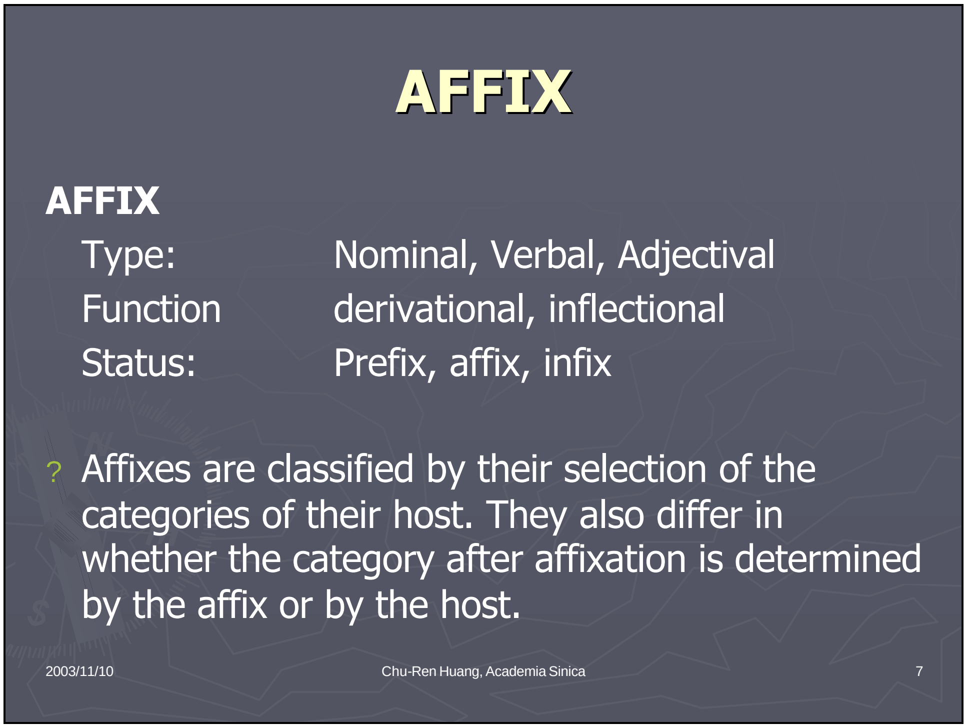
Implications for MILE

? Defining Lemma: How lemmas defined must be clearly stated for each language.

? Languages may have word+affix or stem+affix as their lexical lemmas

? The Category of Affixes

2003/11/10 Chu-Ren Huang, Academia Sinica 6

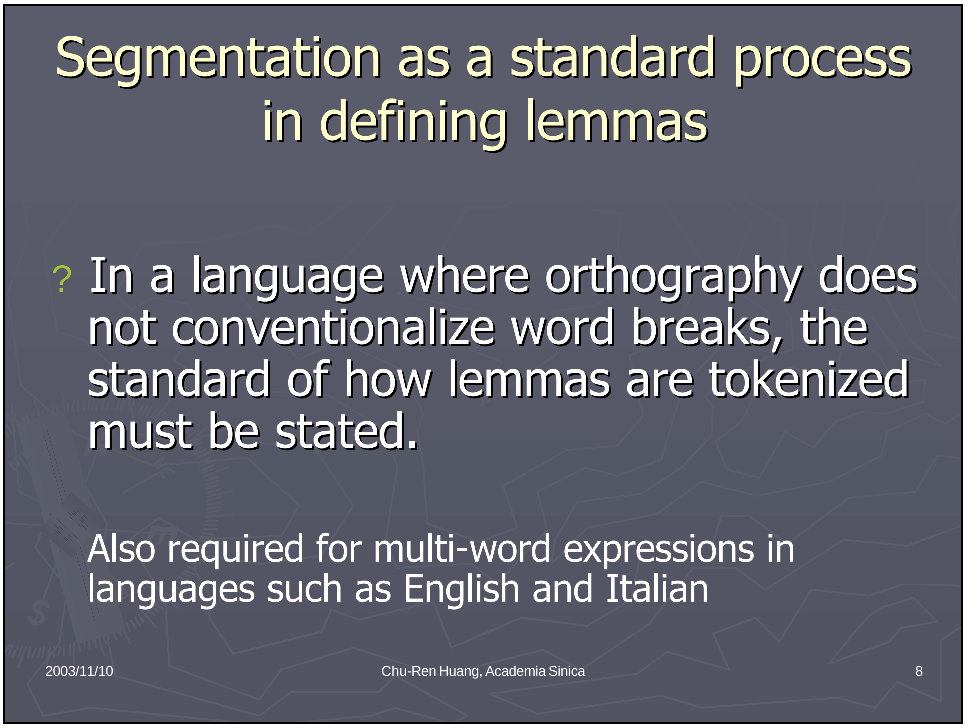


AFFIX

AFFIX

Type:	Nominal, Verbal, Adjectival
Function	derivational, inflectional
Status:	Prefix, affix, infix

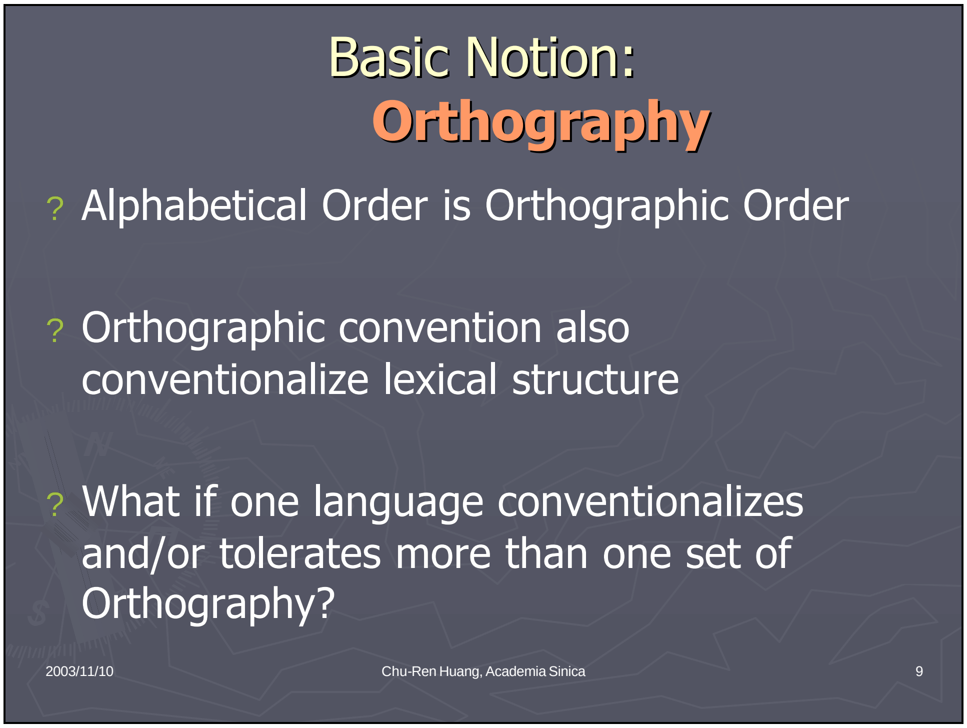
? Affixes are classified by their selection of the categories of their host. They also differ in whether the category after affixation is determined by the affix or by the host.



Segmentation as a standard process in defining lemmas

? In a language where orthography does not conventionalize word breaks, the standard of how lemmas are tokenized must be stated.

Also required for multi-word expressions in languages such as English and Italian

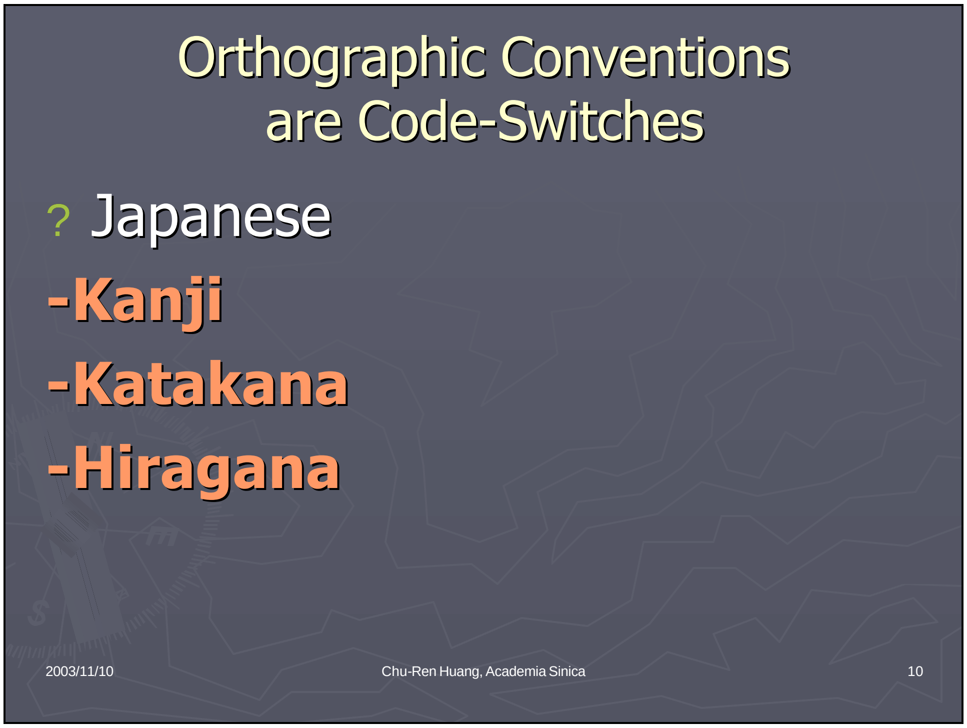


Basic Notion: Orthography

? Alphabetical Order is Orthographic Order

? Orthographic convention also conventionalize lexical structure

? What if one language conventionalizes and/or tolerates more than one set of Orthography?



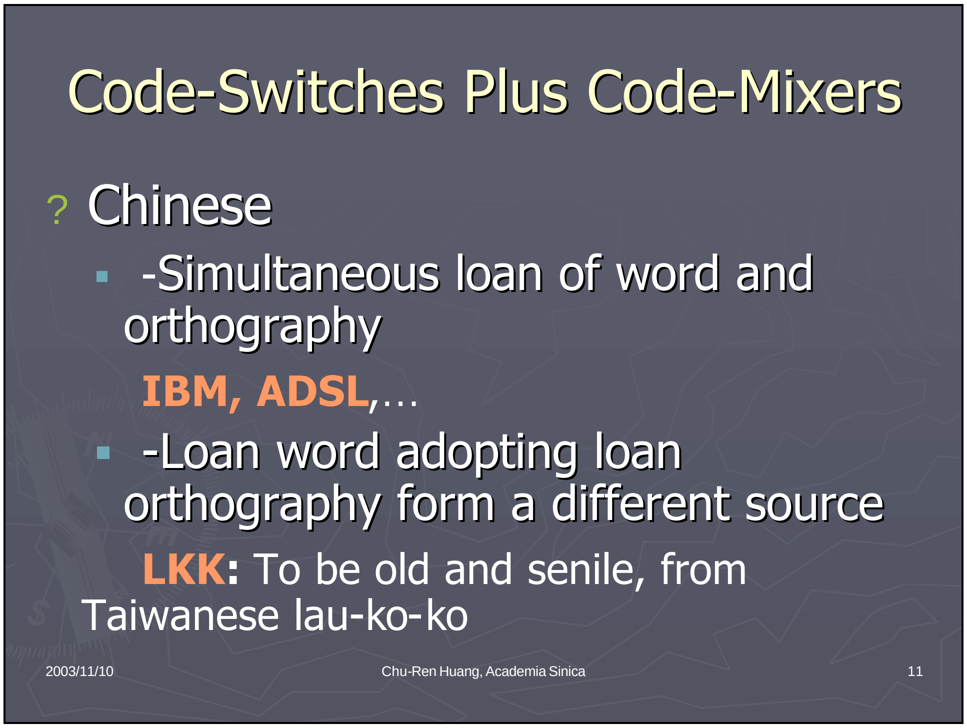
Orthographic Conventions are Code-Switches

? Japanese

-Kanji

-Katakana

-Hiragana



Code-Switches Plus Code-Mixers

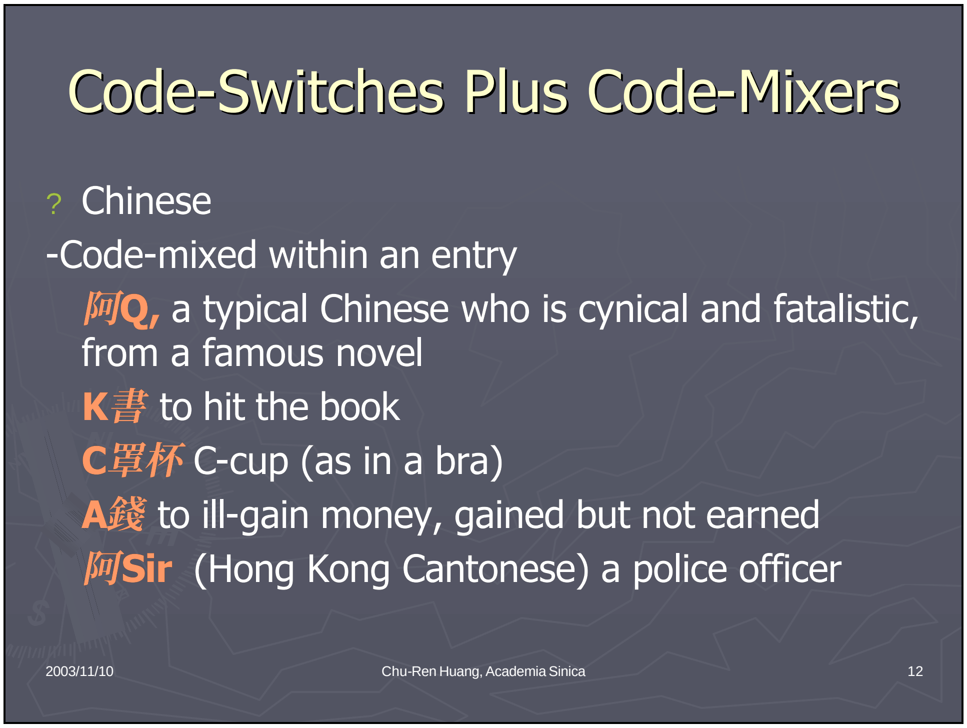
? Chinese

- -Simultaneous loan of word and orthography

  IBM, ADSL,…

- -Loan word adopting loan orthography form a different source

  LKK: To be old and senile, from Taiwanese lau-ko-ko



Code-Switches Plus Code-Mixers

? Chinese

-Code-mixed within an entry

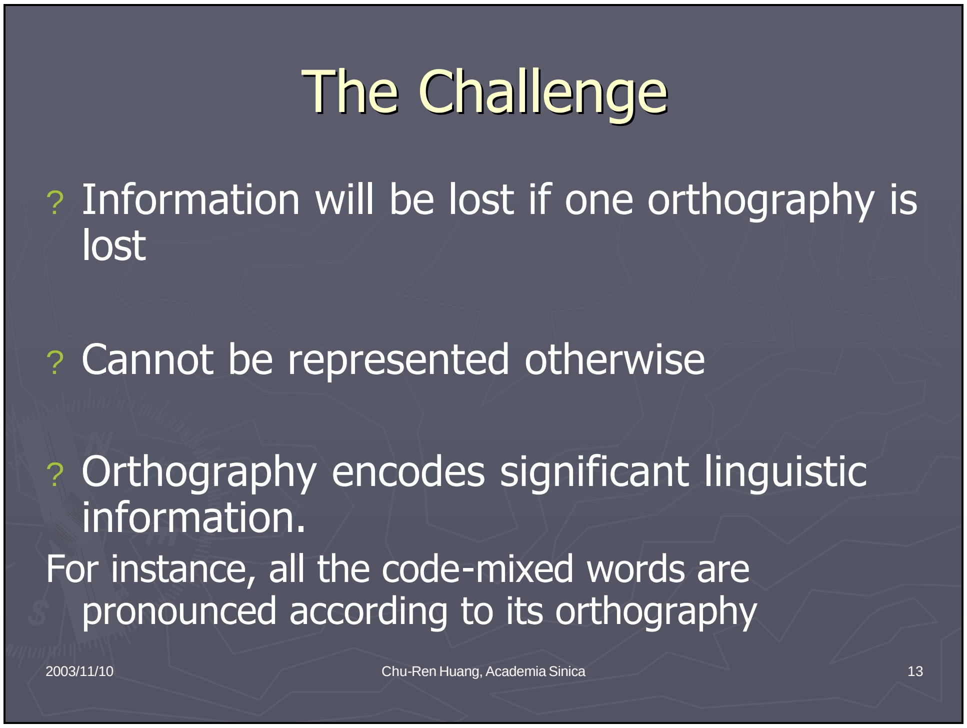
***阿*Q,** a typical Chinese who is cynical and fatalistic, from a famous novel

K書 to hit the book

C罩杯 C-cup (as in a bra)

A錢 to ill-gain money, gained but not earned

***阿*Sir** (Hong Kong Cantonese) a police officer



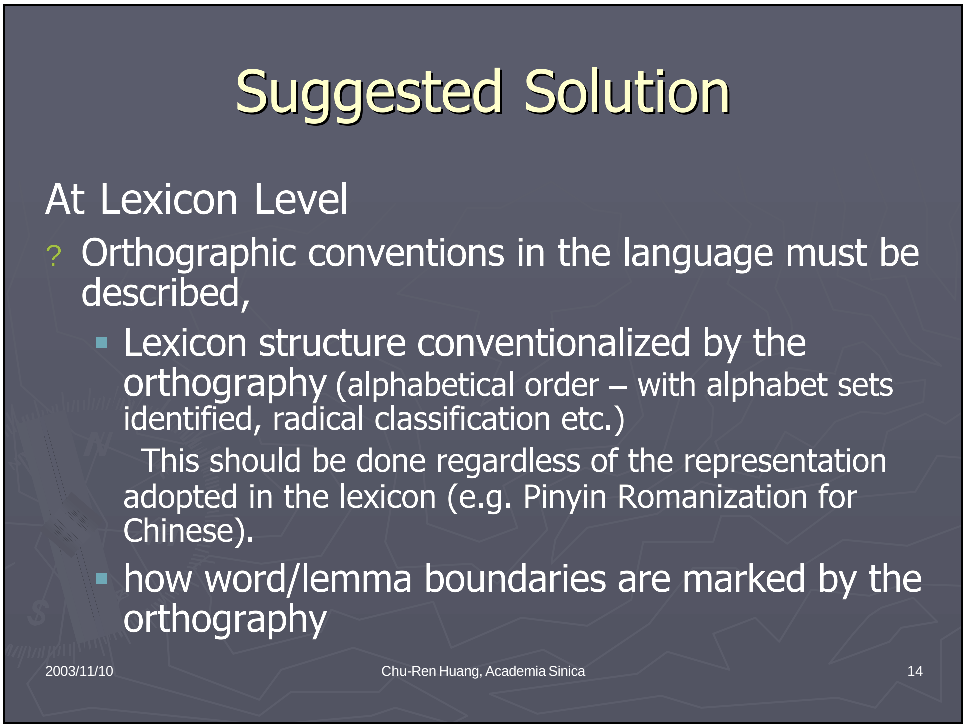
The Challenge

? Information will be lost if one orthography is lost

? Cannot be represented otherwise

? Orthography encodes significant linguistic information.

For instance, all the code-mixed words are pronounced according to its orthography



Suggested Solution

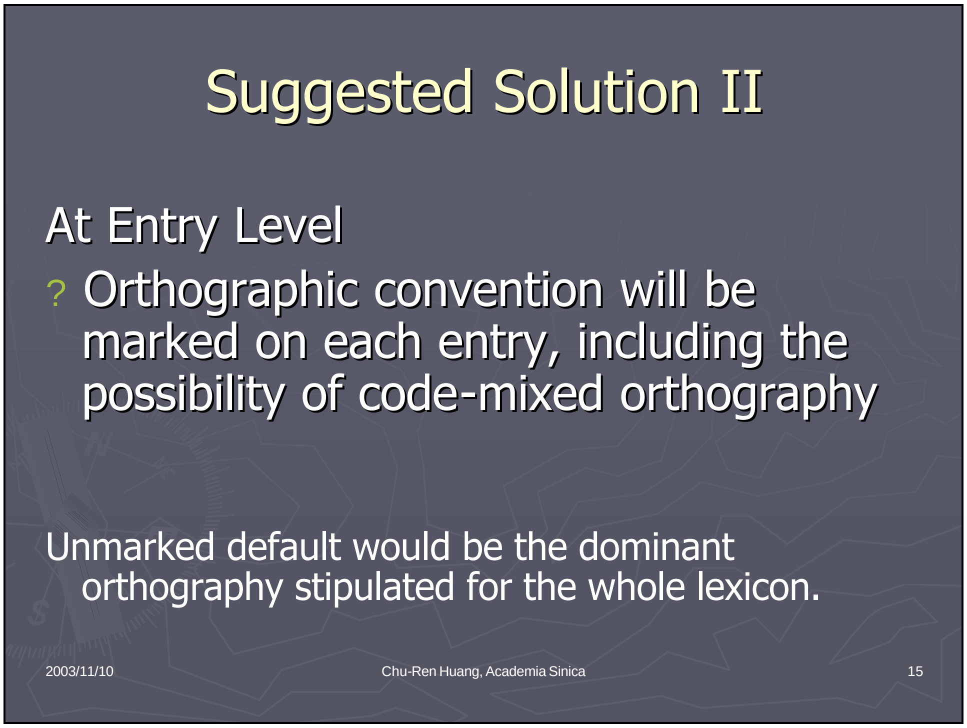
At Lexicon Level

? Orthographic conventions in the language must be described,

- Lexicon structure conventionalized by the orthography (alphabetical order – with alphabet sets identified, radical classification etc.)

  This should be done regardless of the representation adopted in the lexicon (e.g. Pinyin Romanization for Chinese).

- how word/lemma boundaries are marked by the orthography

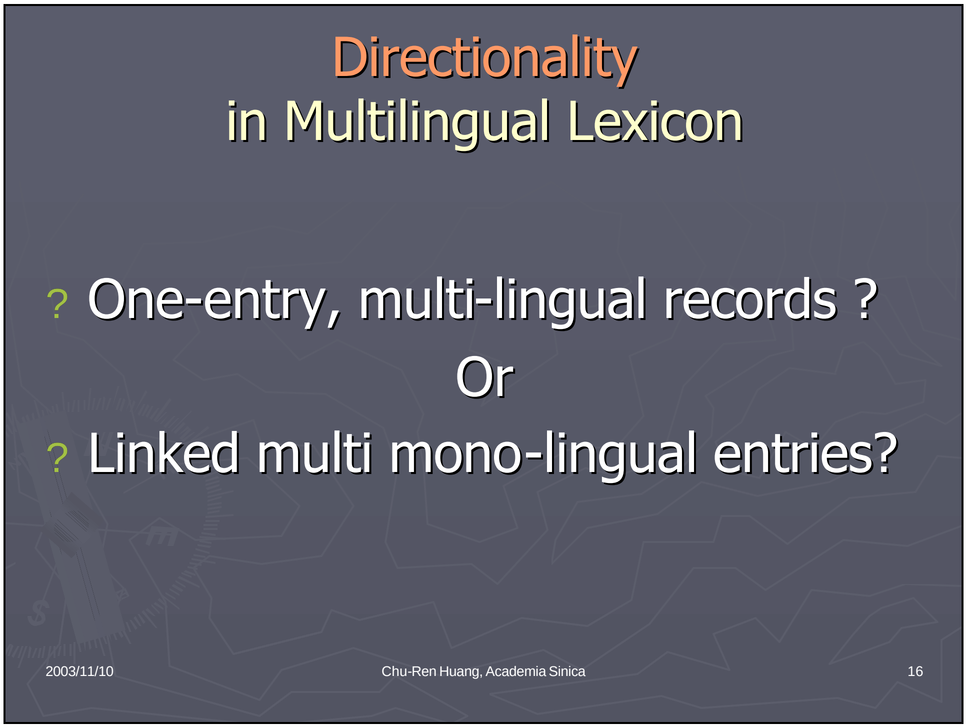


Suggested Solution II

At Entry Level

? Orthographic convention will be marked on each entry, including the possibility of code-mixed orthography

Unmarked default would be the dominant orthography stipulated for the whole lexicon.

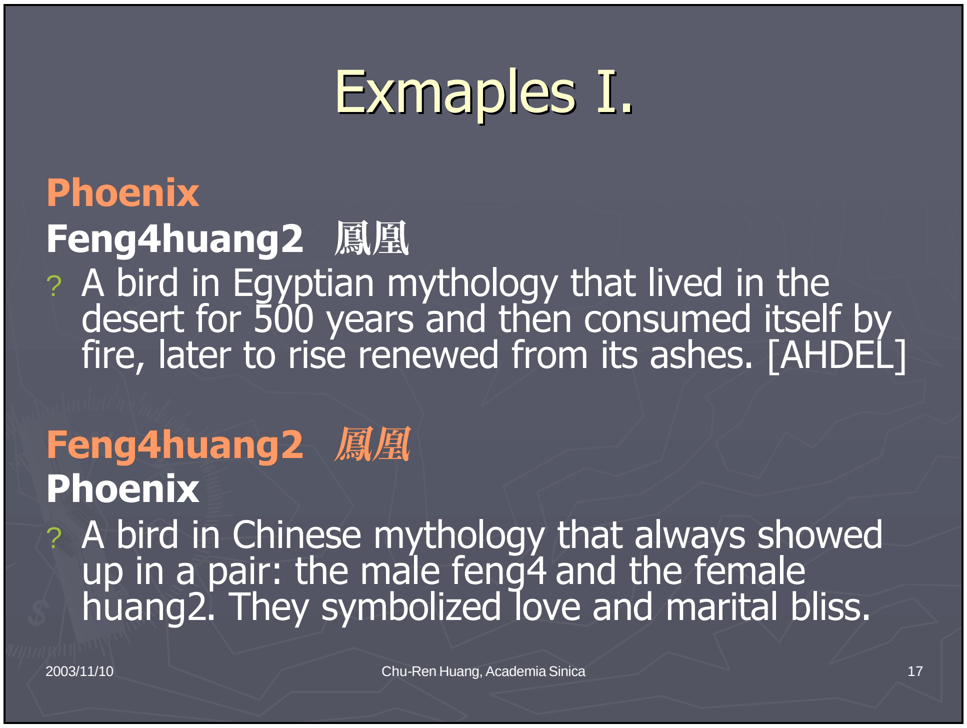


Directionality in Multilingual Lexicon

? One-entry, multi-lingual records ?

Or

? Linked multi mono-lingual entries?



Exmaples I.

Phoenix

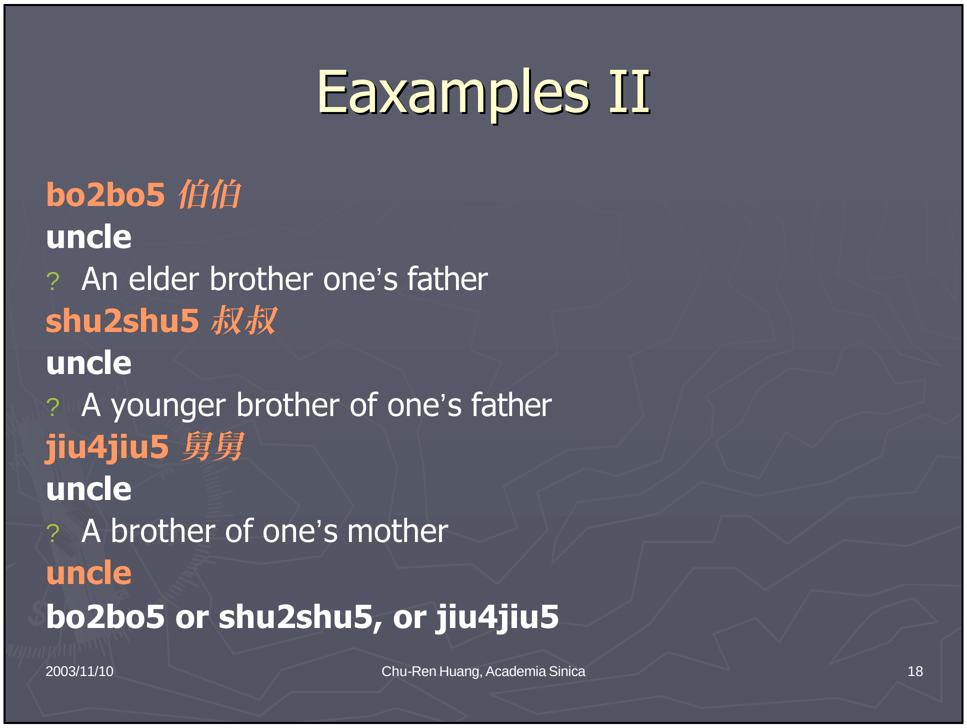
Feng4huang2 鳳凰

- ? A bird in Egyptian mythology that lived in the desert for 500 years and then consumed itself by fire, later to rise renewed from its ashes. [AHDEL]

Feng4huang2 鳳凰

Phoenix

- ? A bird in Chinese mythology that always showed up in a pair: the male feng4 and the female huang2. They symbolized love and marital bliss.



Eaxamples II

bo2bo5 伯伯

uncle

- ? An elder brother one's father

shu2shu5 叔叔

uncle

- ? A younger brother of one's father

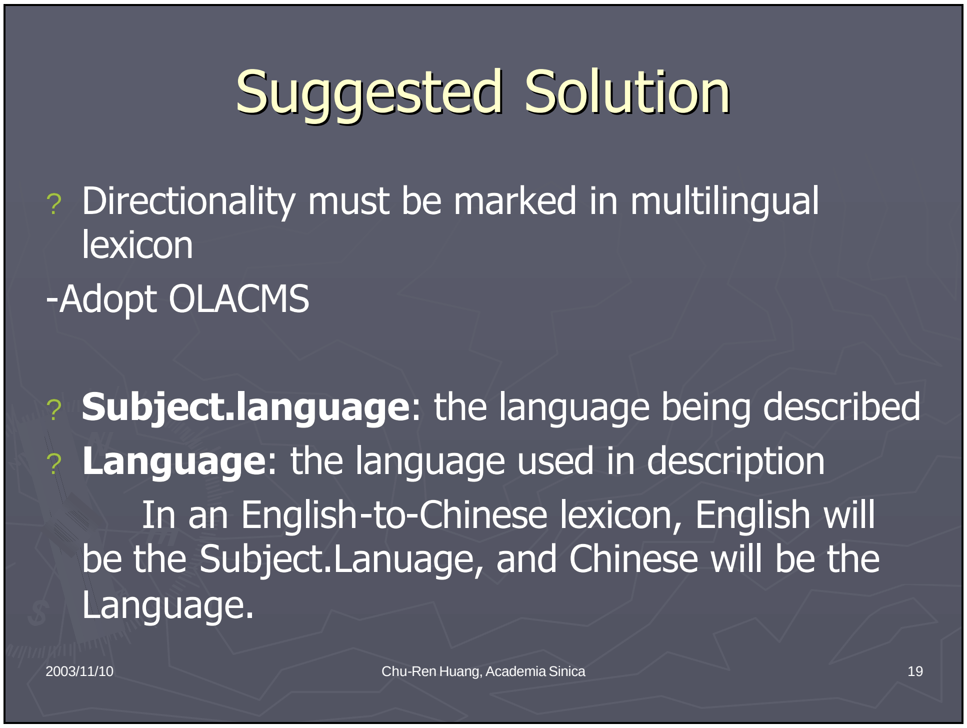
jiu4jiu5 舅舅

uncle

- ? A brother of one's mother

uncle

bo2bo5 or shu2shu5, or jiu4jiu5



Suggested Solution

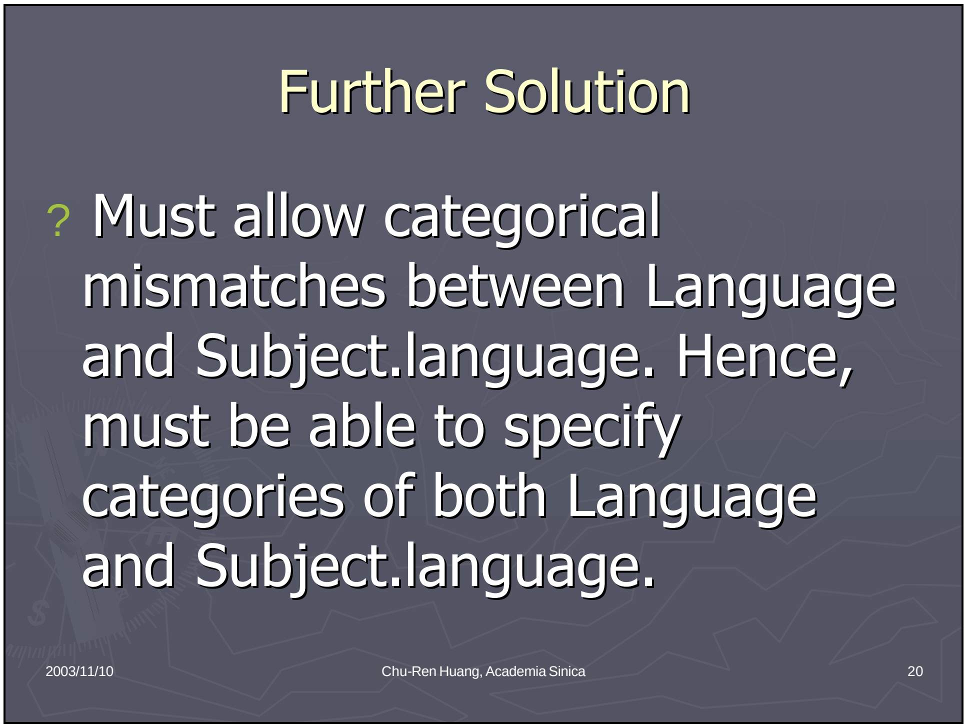
? Directionality must be marked in multilingual lexicon

-Adopt OLACMS

? **Subject.language**: the language being described

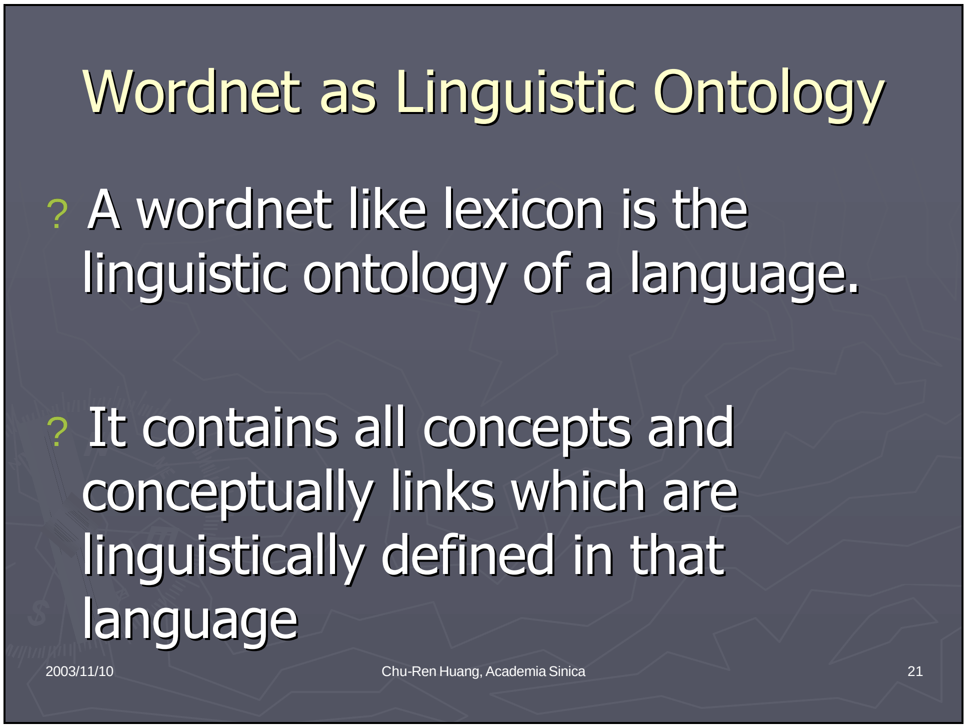
? **Language**: the language used in description

In an English-to-Chinese lexicon, English will be the Subject.Lanuage, and Chinese will be the Language.



Further Solution

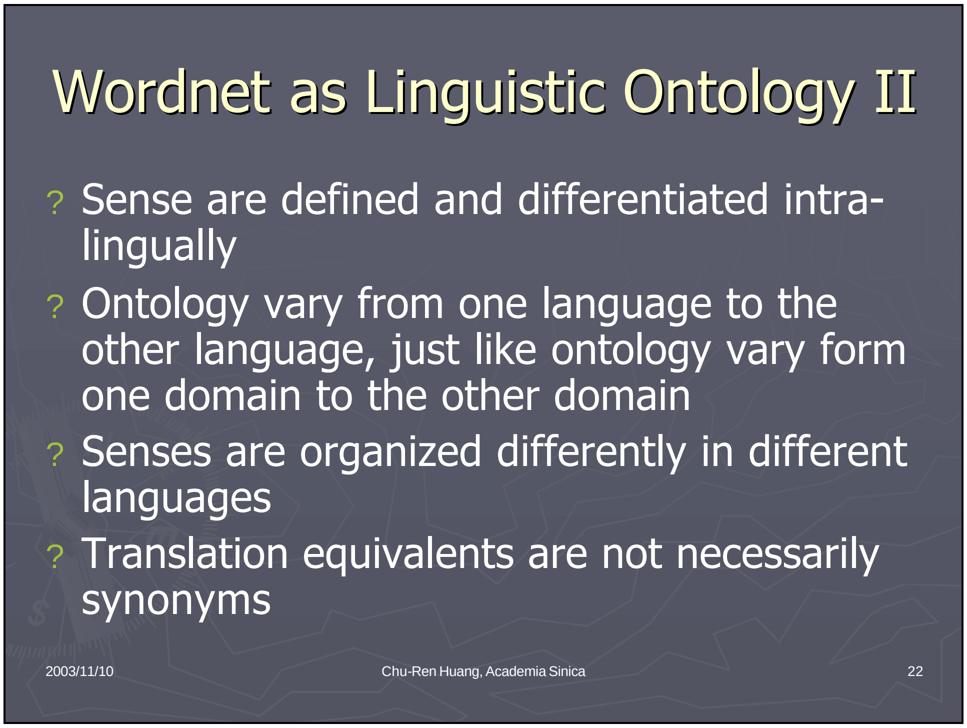
? Must allow categorical mismatches between Language and Subject.language. Hence, must be able to specify categories of both Language and Subject.language.



Wordnet as Linguistic Ontology

? A wordnet like lexicon is the linguistic ontology of a language.

? It contains all concepts and conceptually links which are linguistically defined in that language



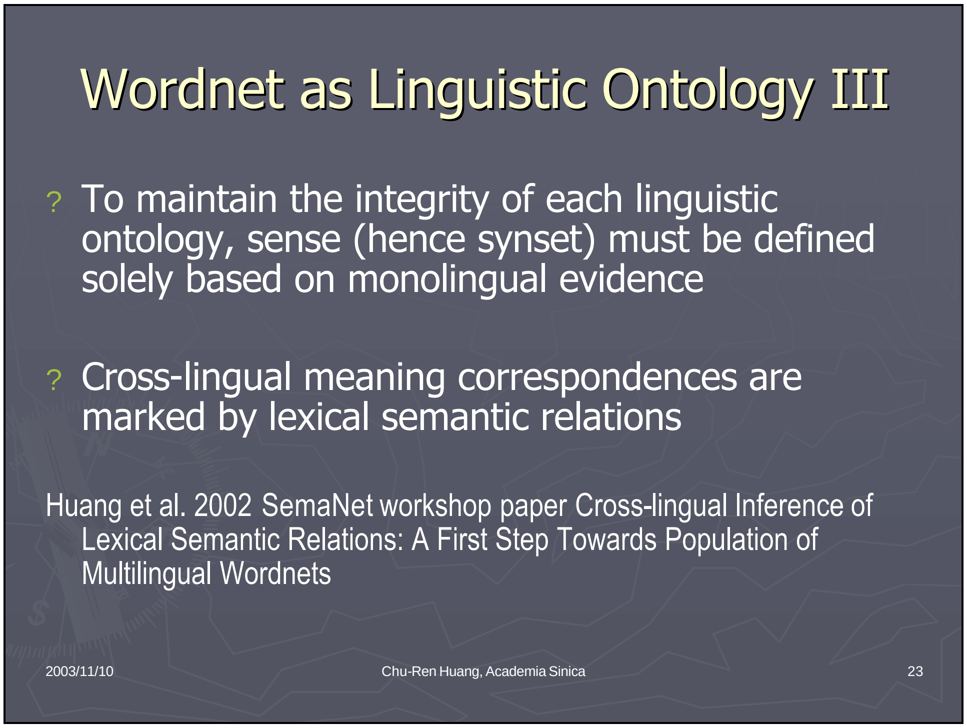
Wordnet as Linguistic Ontology II

? Sense are defined and differentiated intra-lingually

? Ontology vary from one language to the other language, just like ontology vary form one domain to the other domain

? Senses are organized differently in different languages

? Translation equivalents are not necessarily synonyms

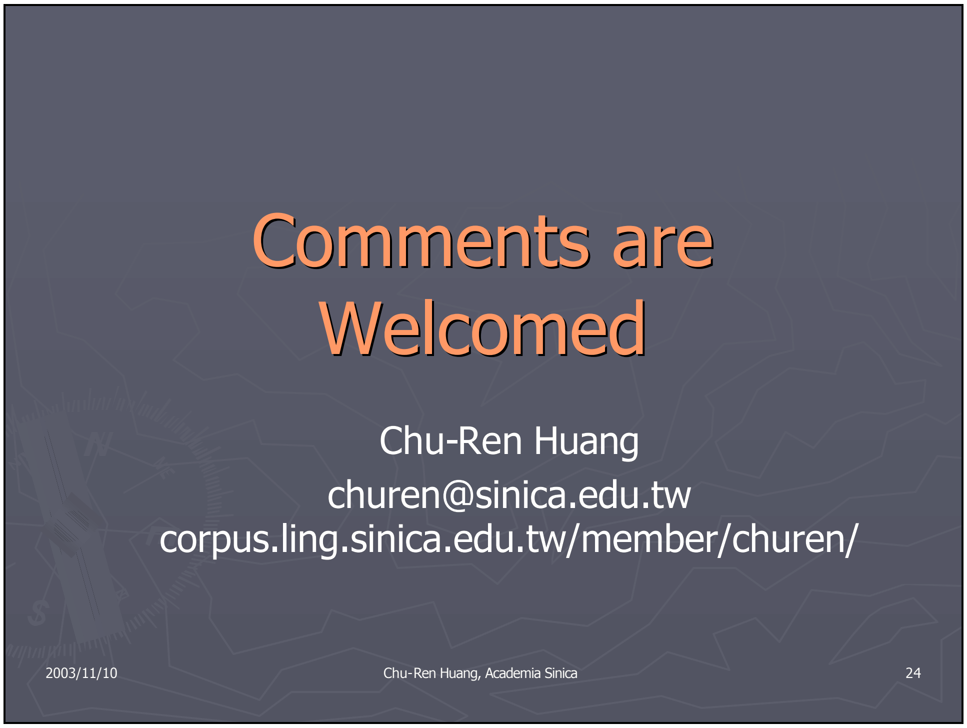


Wordnet as Linguistic Ontology III

- ? To maintain the integrity of each linguistic ontology, sense (hence synset) must be defined solely based on monolingual evidence

- ? Cross-lingual meaning correspondences are marked by lexical semantic relations

Huang et al. 2002 SemaNet workshop paper Cross-lingual Inference of Lexical Semantic Relations: A First Step Towards Population of Multilingual Wordnets



Comments are Welcomed

Chu-Ren Huang

churen@sinica.edu.tw

corpus.ling.sinica.edu.tw/member/churen/