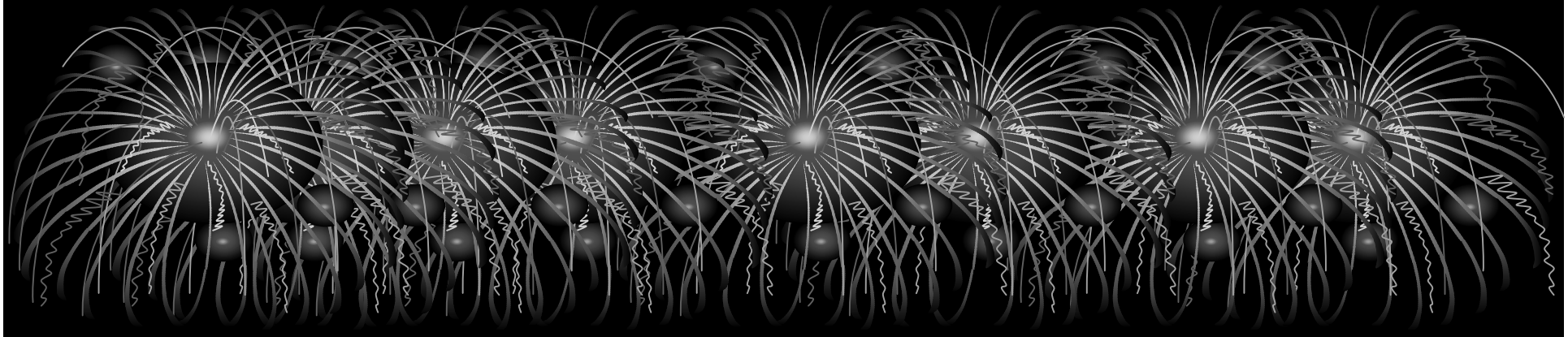


A Cross System Machine Translation

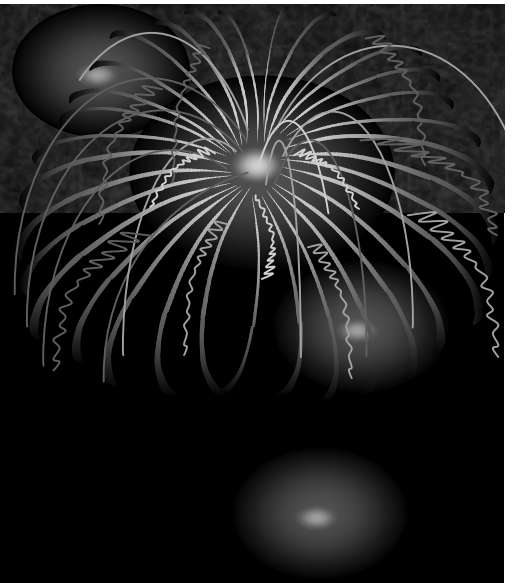
Thepchai Supnithi

Virach Sornlertlamvanich

Thatsanee Chareonporn

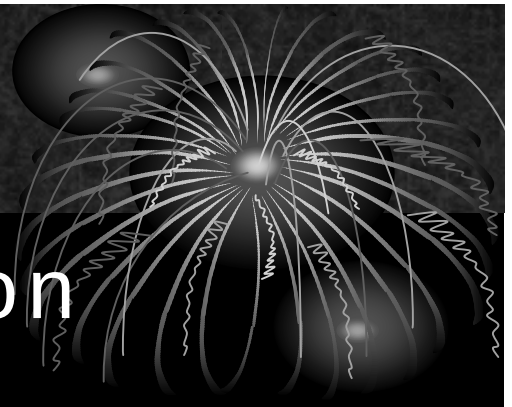
National Electronics and Computer TEchnology Center(NECTEC)

Outline



- Background
- Problem Identification
- Cross System Technology
- Linguistic Problems
- Example of Information Transfer in Cross System MT
- Conclusion and Future Work

Background

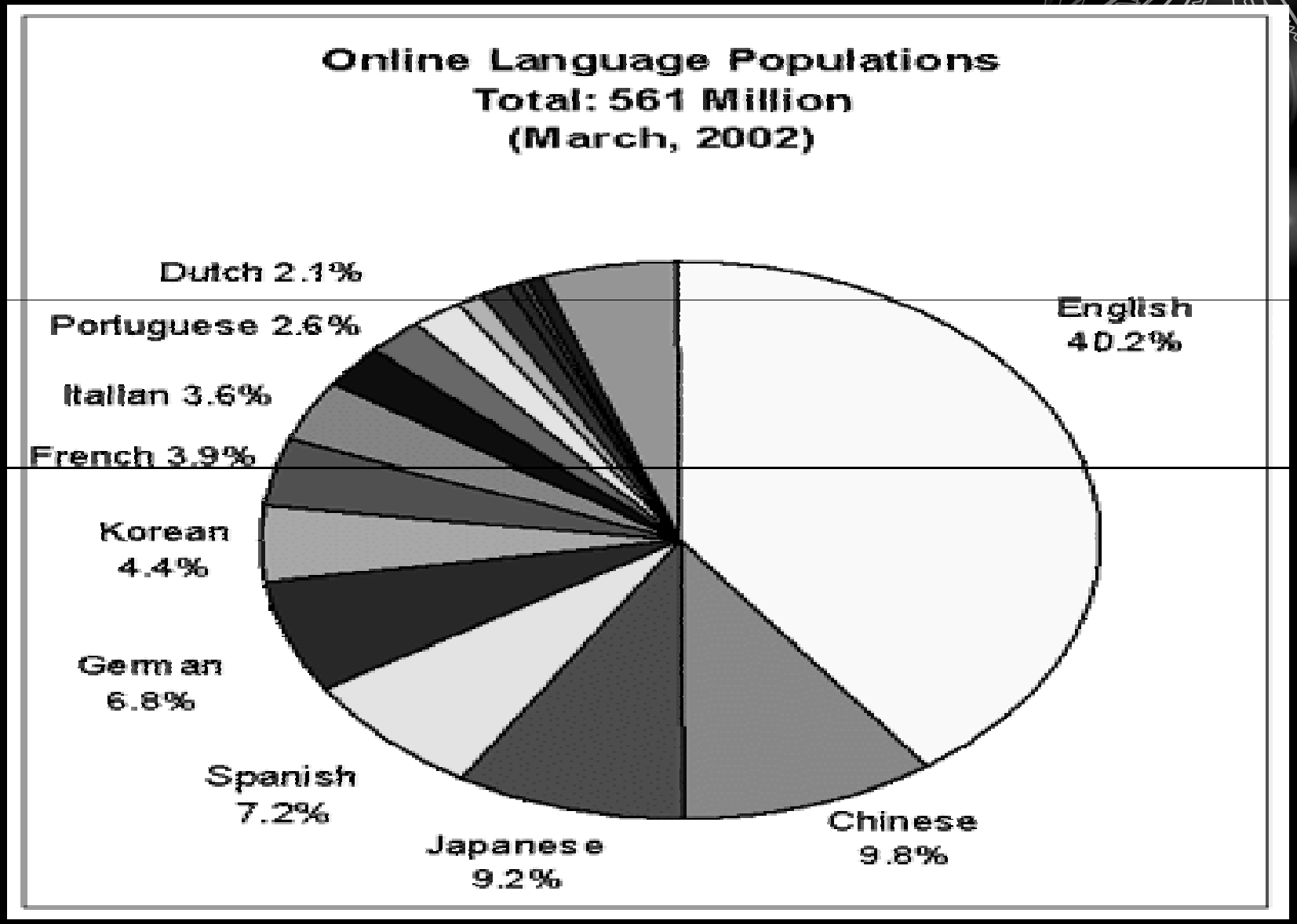


- Rapidly Growth of Information
  - Digital Divide Problem

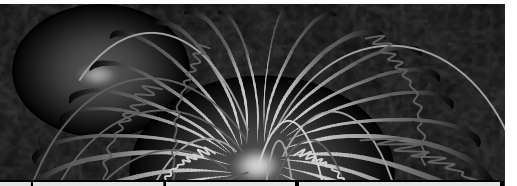
    Digital divide : the divide between those with access to new technologies (information) and those without

- Cost and time consumption for developing
  - Small countries in which there are not NLP/MT fundamental research.
  - Find a good way to help them build only essential resources/tools

Global Internet Statistics (by Language)



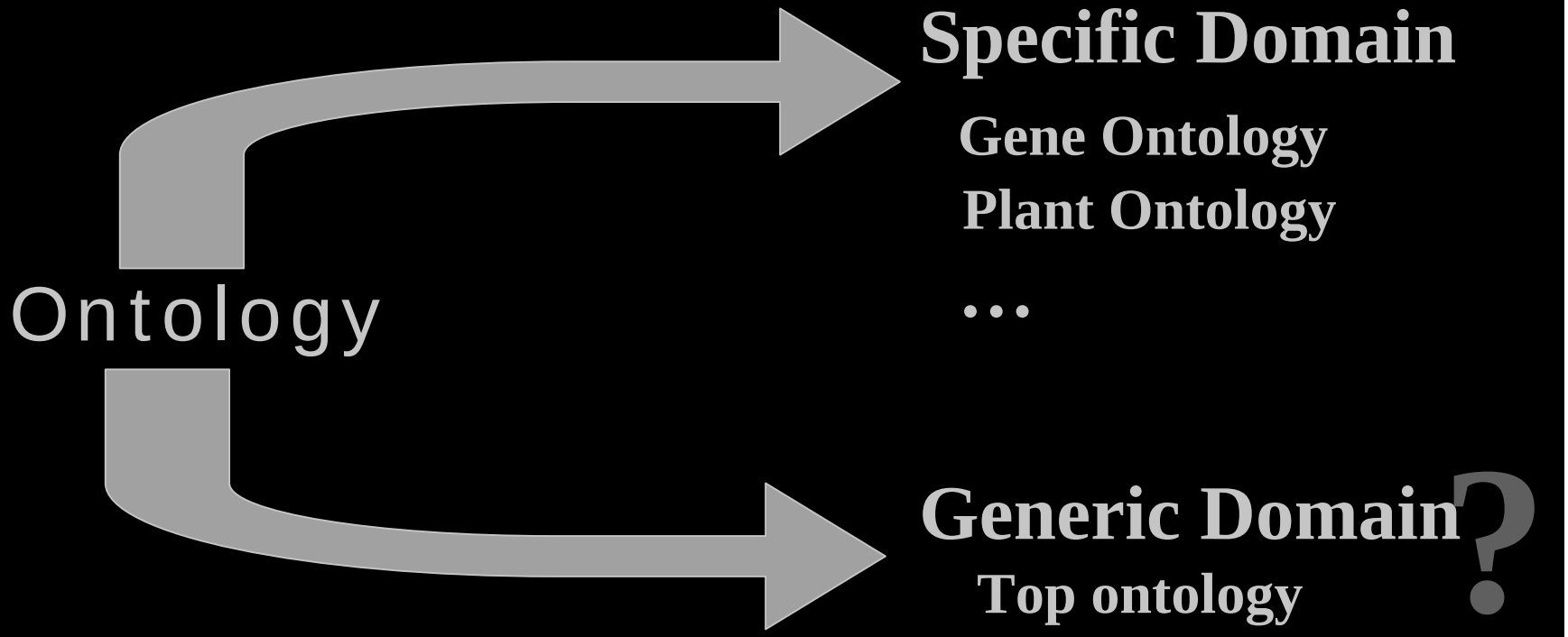
Online Language in Asia



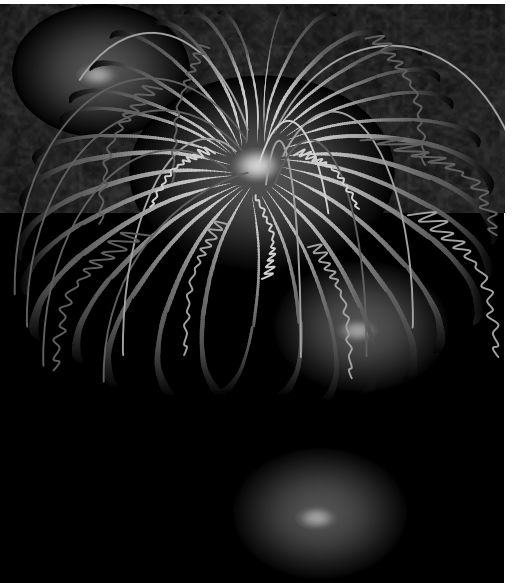
	Chines	Japane	Korea	Engli	Malay	Tha	Arab	Hebr	Persi	Serbi	Vietn	TOTAL
TOTAL:	53.103	51.288	24.1	9.63	4.807	2.3	2.036	1.93	1.3	0.3	0.1	150.894
Bahrain							0.105					0.105
China	33.700											33.700
Hong Kon	4.310											4.310
India				6.000								6.000
Indonesia					2.000							2.000
Iran									1.300			1.300
Iraq							0.013					0.013
Israel								1.930				1.930
Japan		51.288										51.288
Jordan							0.088					0.088
Korea			24.100									24.100
Kuwait							0.165					0.165
Lebanon							0.263					0.263
Malaysia	1.233			0.200	2.467							3.900
Oman							0.084					0.084
Pakistan				1.200								1.200
Philippines				2.000								2.000
Qatar							0.075					0.075
Saudi Arabia							0.570					0.570
Serbia										0.300		0.300
Singapore	2.260			0.230	0.340							2.830
Taiwan	11.600											11.600
Thailand						2.300						2.300
U.A.E							0.660					0.660
Vietnam											0.100	0.100
Yemen							0.014					0.014

Problem Identification for Multilingual MT (1)

- Difficulty of Building Interlingua
  - Interlingua is an ideal language.
    - Difficulty in completing all concepts



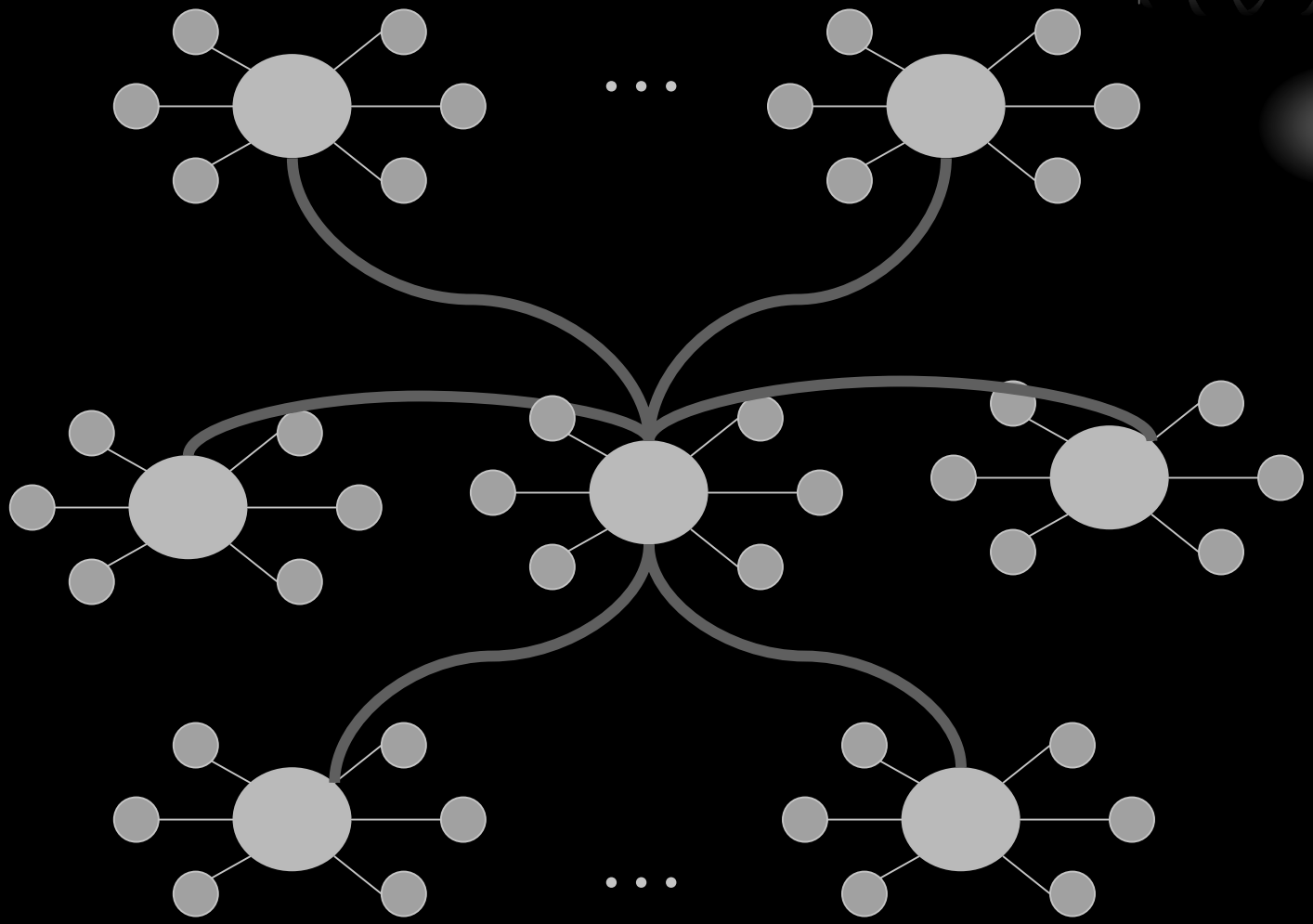
Problem Identification for Multilingual MT (2)



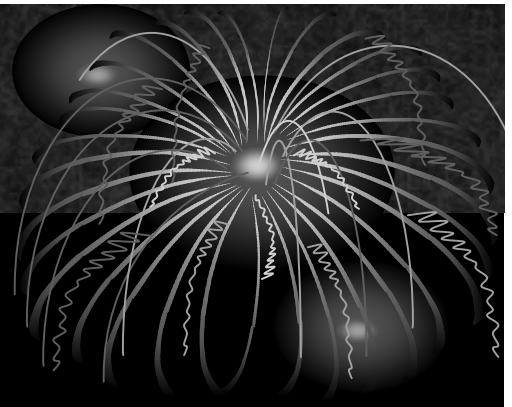
- Languages Dependency
  - Same language family
    - French-English
  - Difference language family
    - Thai-English

Cross System Idea

- Connecting among language families

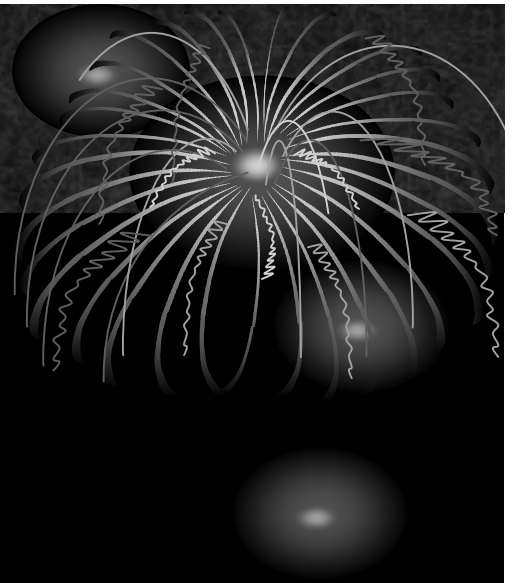


Cross System approach



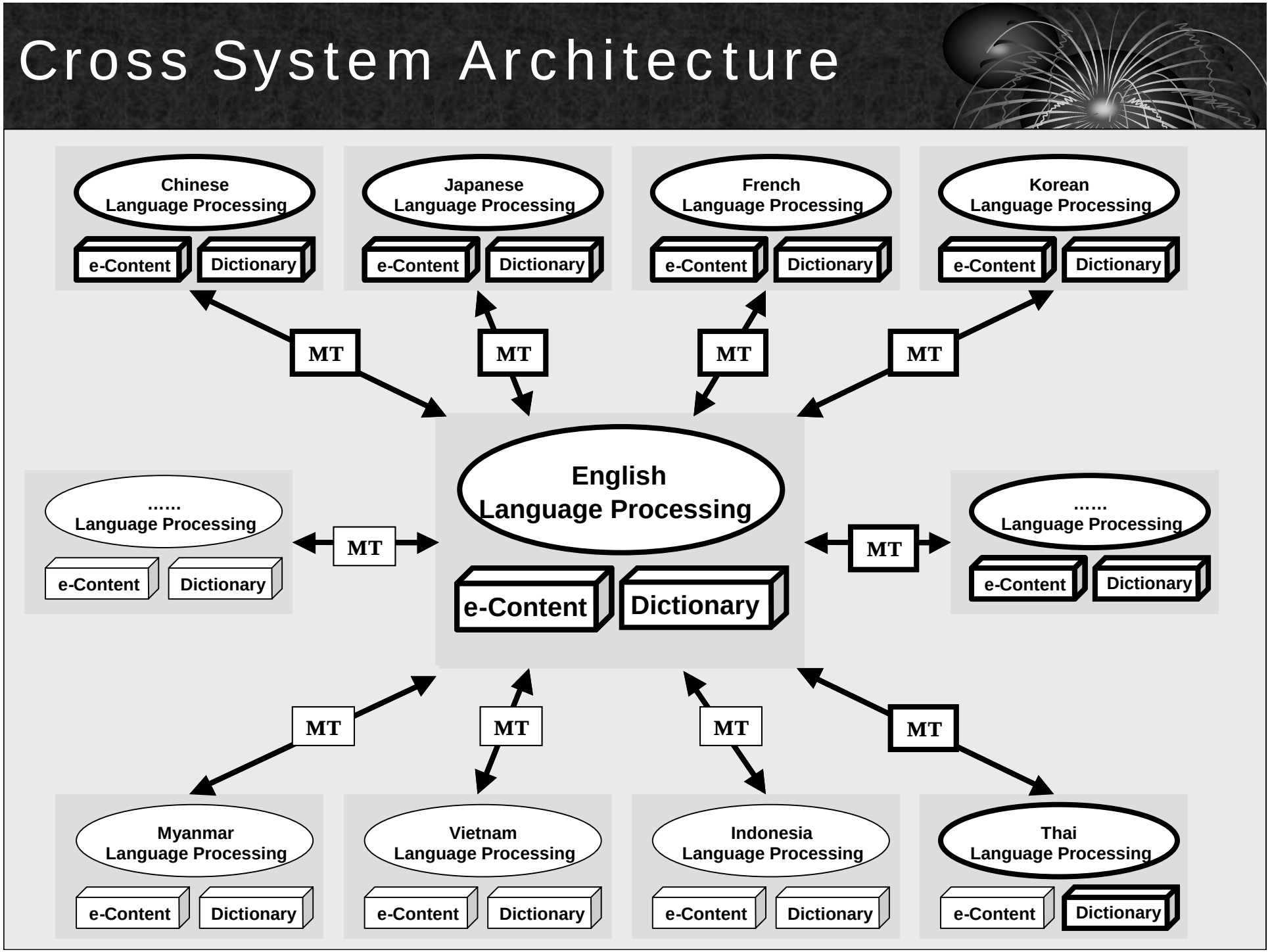
- Resources Reuse
  - A lot of MT systems are developed
  - The 20 most spoken language in the world are also have their own MT systems.
- Language Sharing
  - Cross System Intermediate Language
    - English (difference language family)
      - User Viewpoint : The biggest resources
      - Developer Viewpoint :
        - Second most as a communication
        - Easier to connect with other existing resources

Cross System approach



- Distributed System
  - Bilingual MT System
    - Local Language←→ English
  - System Type Independent
    - Possible for all Types of MT System

Cross System Architecture

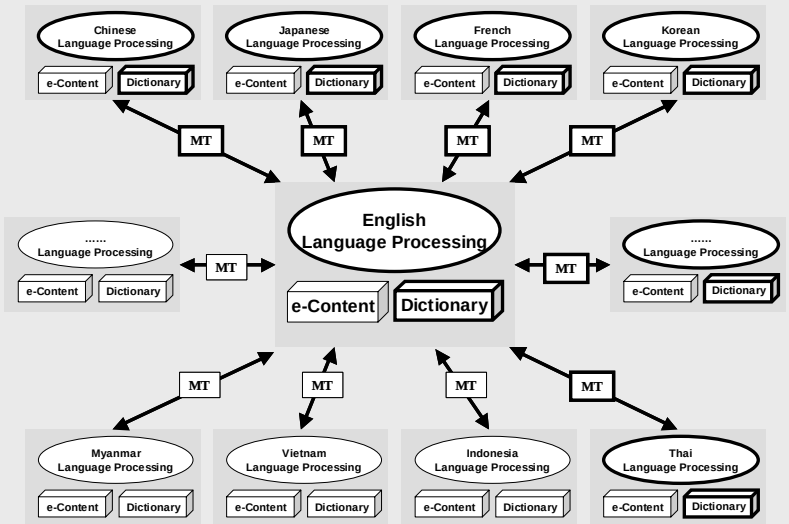


Application Tools

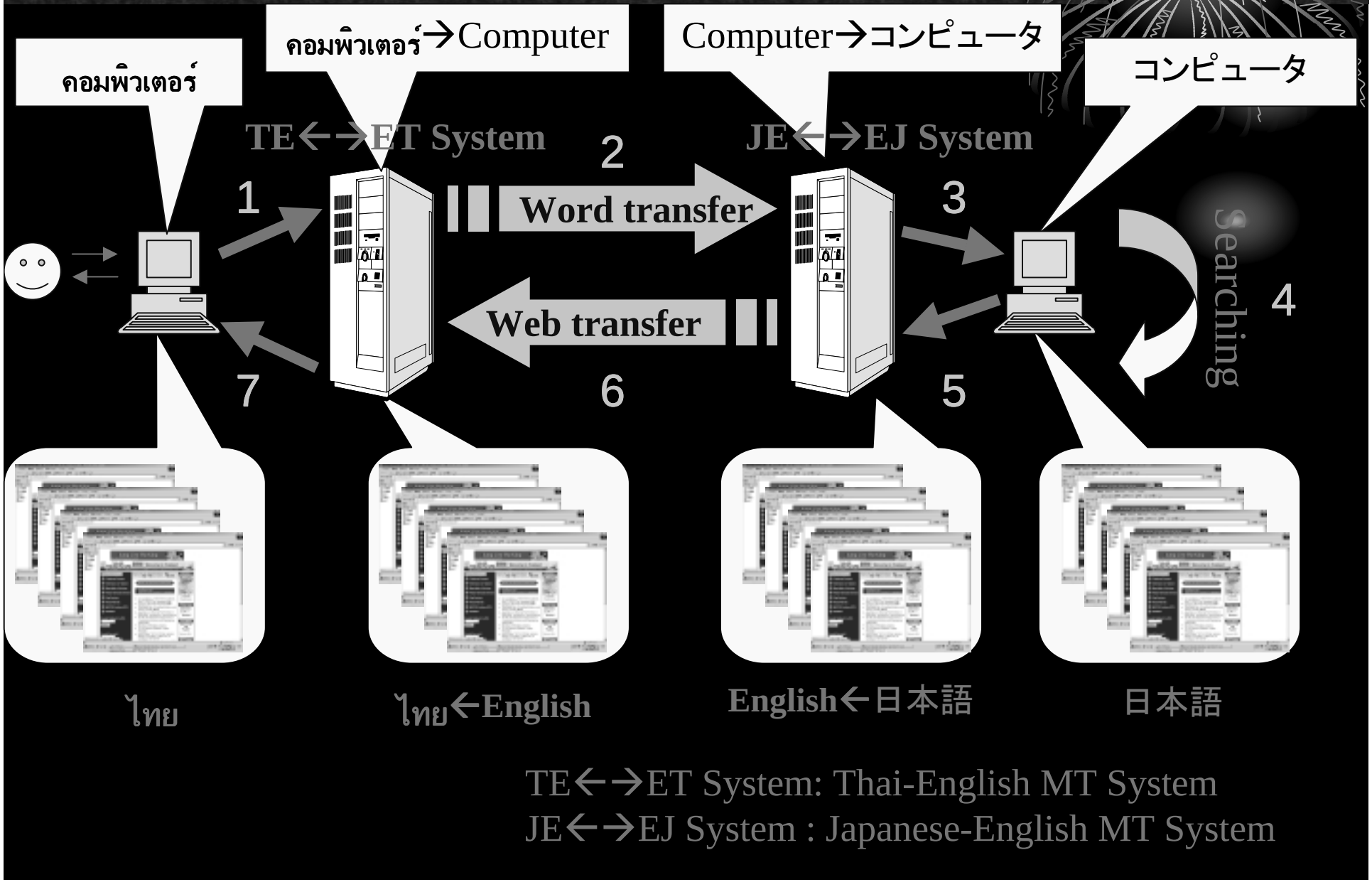
E-services

Representation | Extraction | Retrieval | Summarization | MT | Mining | Visualization

Cross System Technology

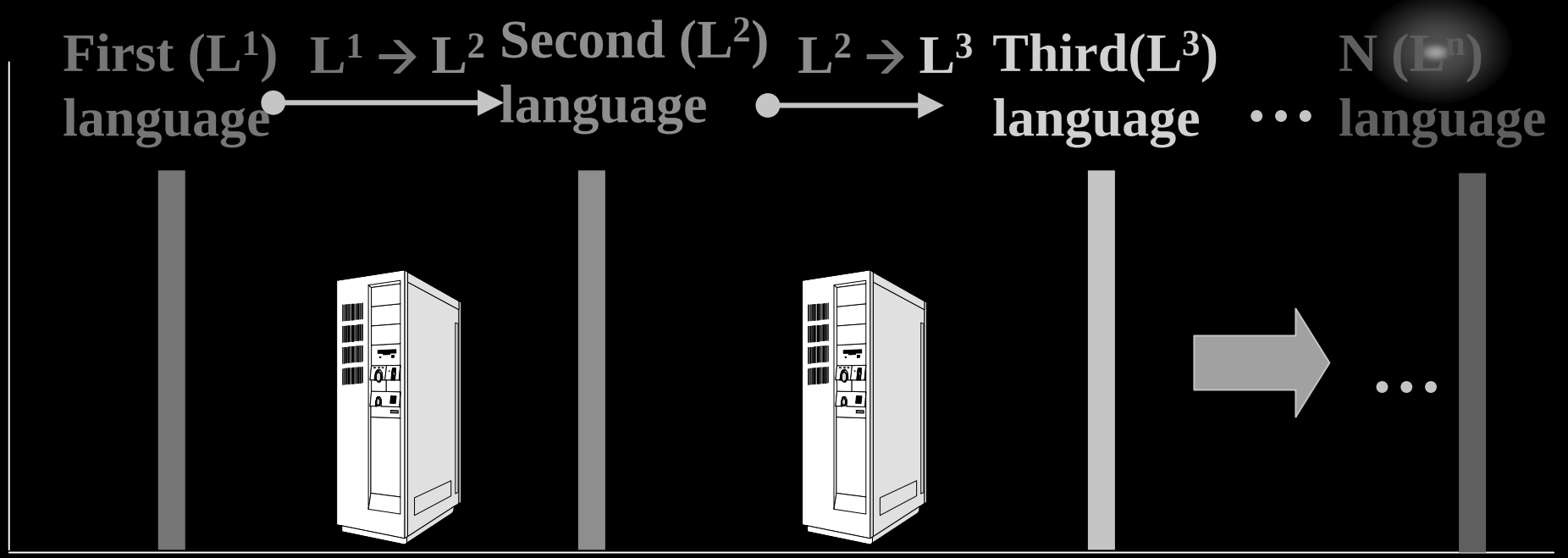


Information Retrieval

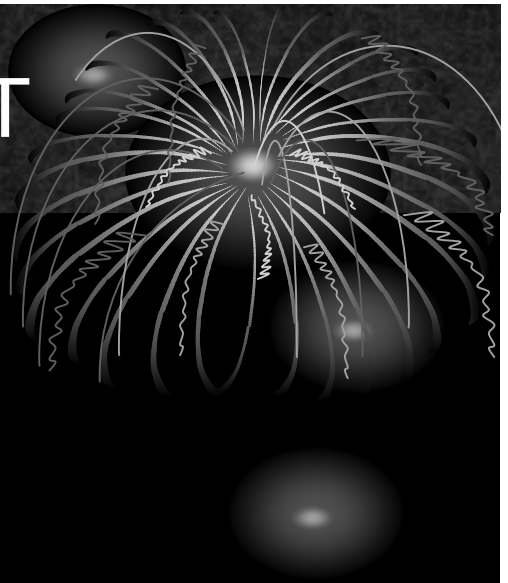


TE←→ET System: Thai-English MT System
JE←→EJ System : Japanese-English MT System

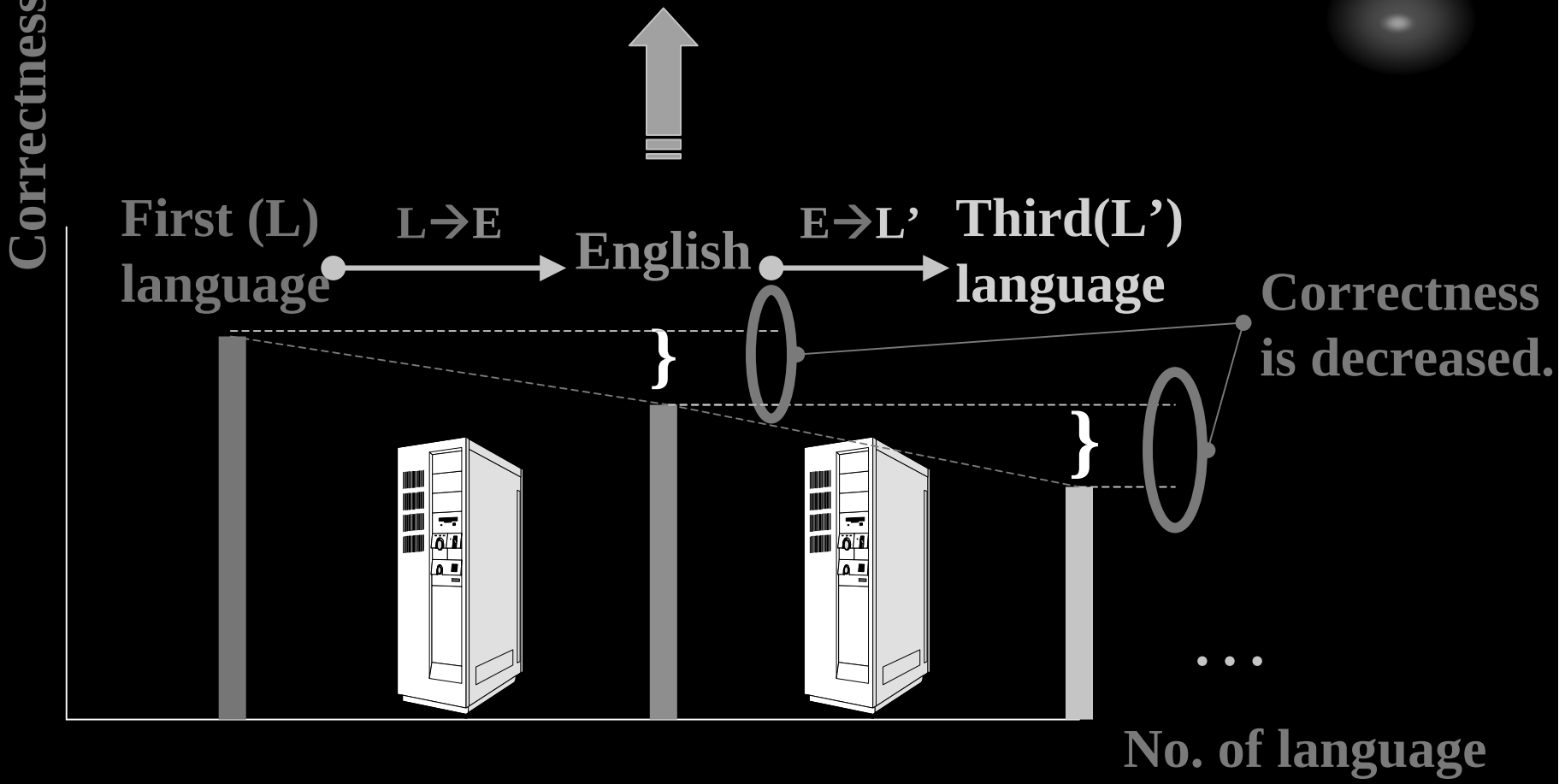
An Ideal Efficiency of Cross System MT



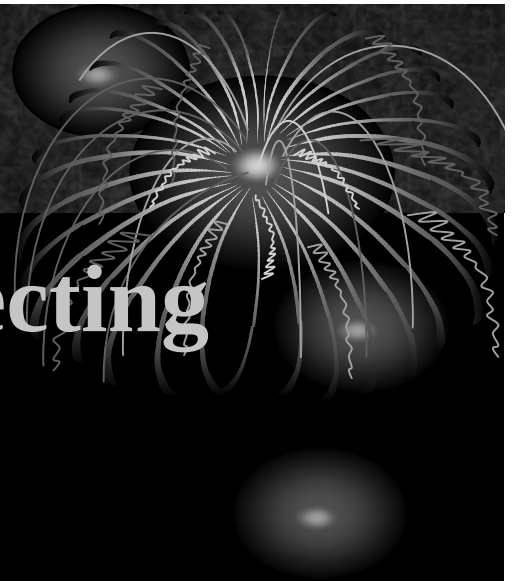
A Problem of Cross System MT



Increasing Correctness:
Keep Original Information



Tag Structure



We use XML format for collecting the original information

- Syntax
- Semantics
- Pragmatics
- Discourse

Sentence Analysis Statistics

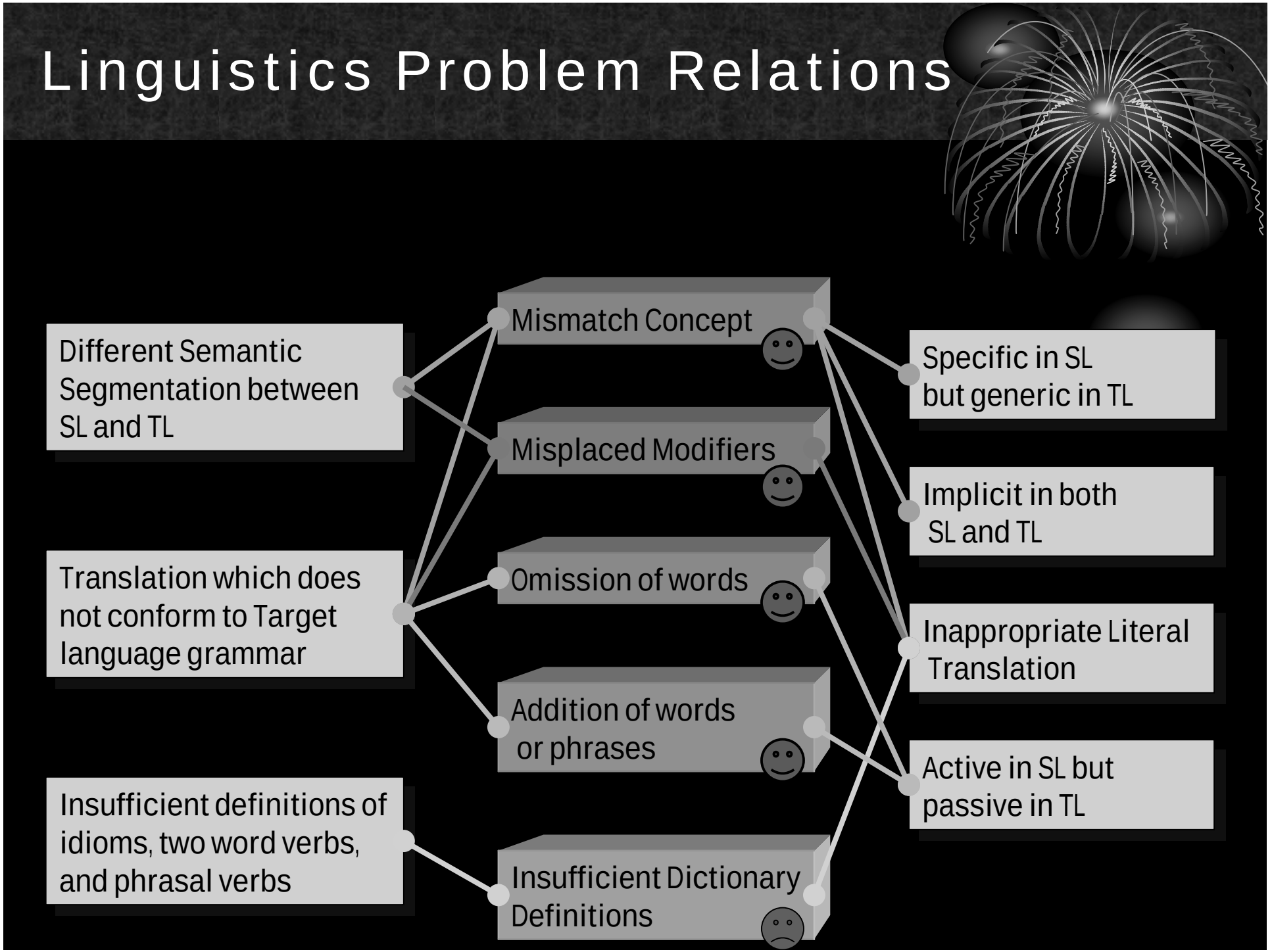
Analysis of 770 sentences from E-T machine translation system

Categories	No. of Sentences	Percent
Perfect Sentence	224	29.09
Comprehensible Sentence	429	55.71
Incomprehensible Sentence	117	15.20
Total	770	100

Survey of Linguistics Problem

Linguistics Problems	Meaning
Mismatch Concept	Inappropriate concept is selected
Misplaced Modifiers	Wrong position of words, phrases or modifiers in TL resulting in distortion of meaning
Inappropriate Literal Translation	An inappropriate translation that follows closely the form of SL. It can be categorized into 1) part of speech, 2) order, 3) idiom
Addition of words or phrases	Some words in TL that are not stated in SL are added
Omission of words	The meaning of a word or words when translating from SL to TL is/are leaved out
Insufficient definitions of idioms, two word verbs, and phrasal verbs	The scope or number of words in electronic storage is either limited or inaccurate according to the meanings of words in SL
Translation which does not conform to Target language grammar	A difference sentence structure in TL that may cause an incomprehensible translation
Implicit in both SL and TL	The implied meaning of a word in the SL is not expressed clearly or fully in TL
Active in SL but passive in TL	The participles appear in SL as active forms but are translated into passive forms in TL
Insufficient Dictionary Definitions	The scope or number of words in the electronic data dictionary is limited
Different Semantic Segmentation between SL and TL	Using difference marker, such as punctuation or space in SL and TL may cause the incomprehensible translation
Specific in SL but generic in TL	A specific word in SL is referred as a general meaning in TL

Linguistics Problem Relations



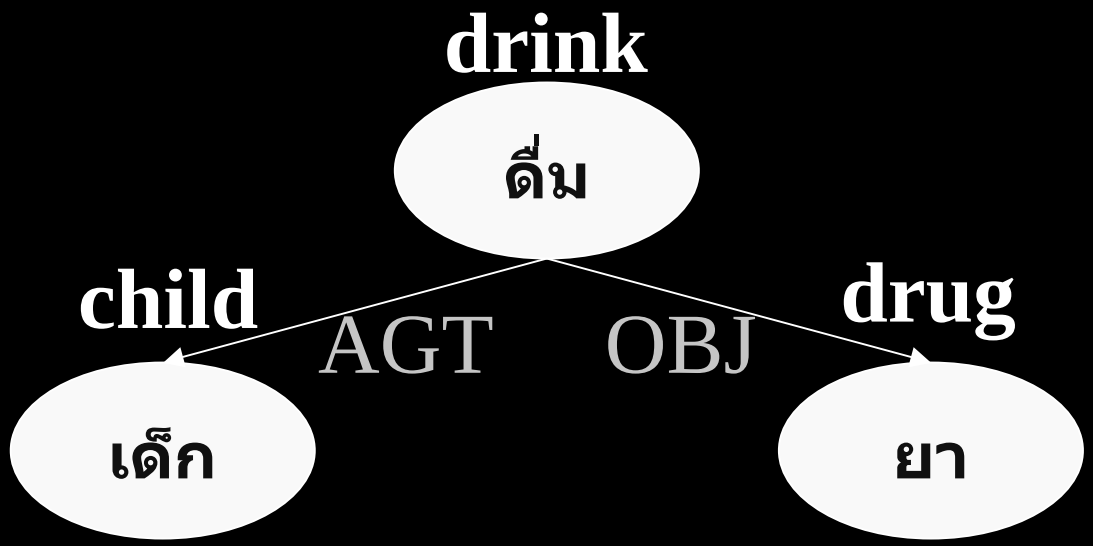
Example for Transfer Information

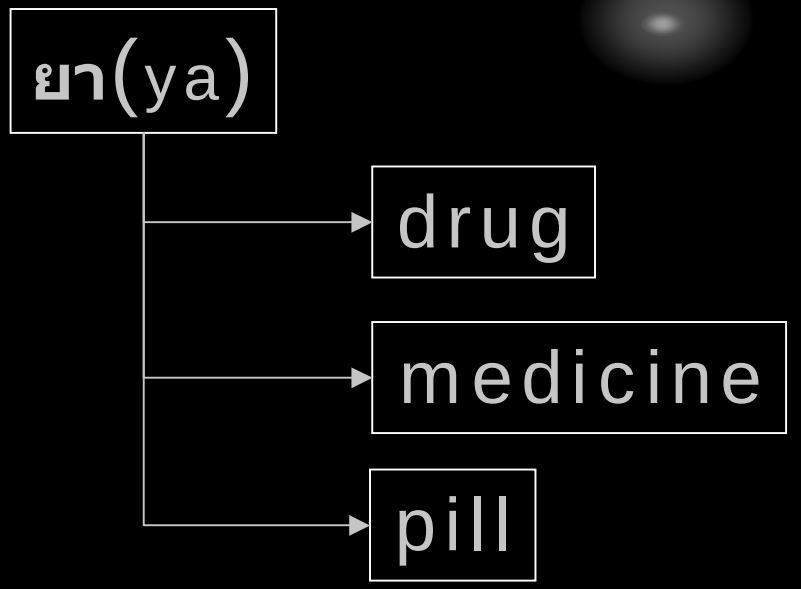
Mismatch Concept

- **Ex 1) เด็กดื่มยา (dek duum ya)**

M A child drinks a drug .

O A child drinks a medicine .





Example for Transfer Information (cont)

- **Ex 1) เด็กดื่มยา (dek duum ya)**

  M A child drinks a drug .

  ○ A child drinks a medicine .

A child <**AGT**> drink a **drug** <**OBJ:** **c# drug,medicine,pill**>.

Example for Transfer Information (cont)

Misplaced Modifier

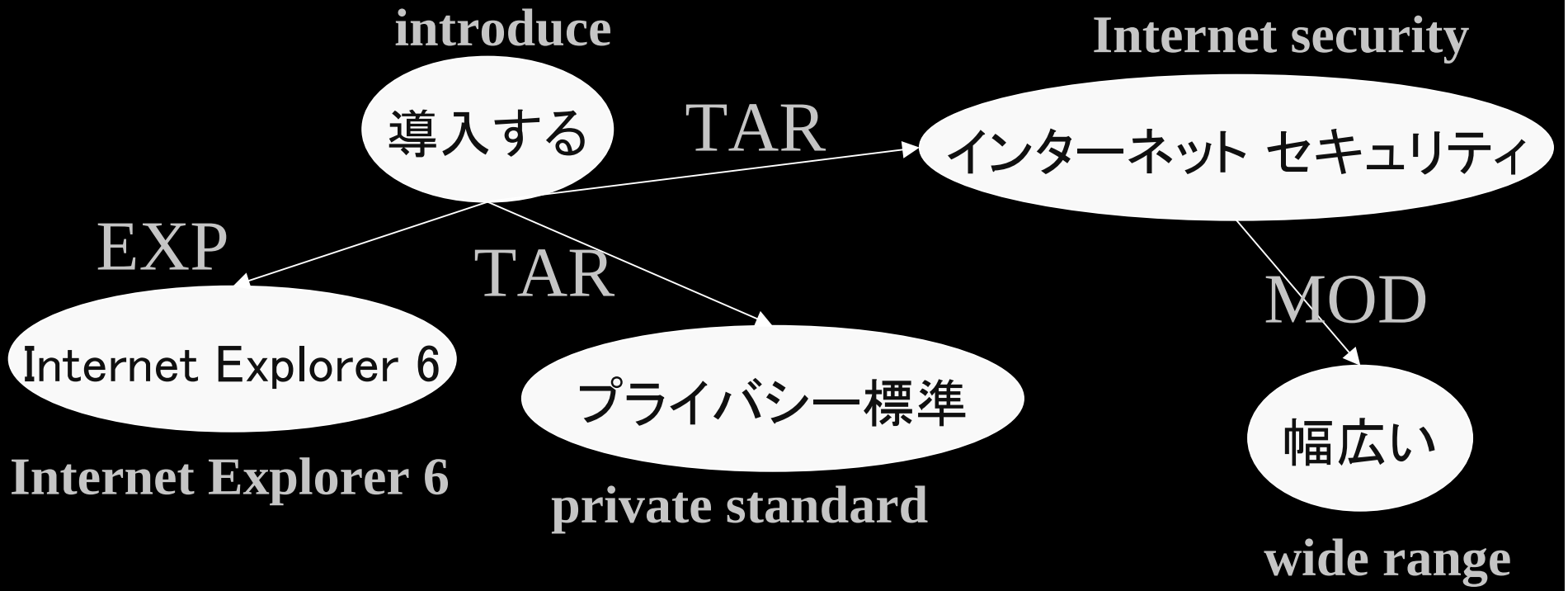
- **Ex2)Internet Explorer 6** では、幅広いインターネット セキュリティとプライバシー標準を導入しています。

M **A private standard has been introduced as the Internet security whose it is wide range in InternetExplorer6.**

○ **A private standard and a wide range internet security has been introduced in InternetExplorer6.**

Example for Transfer Information (cont)

- **Internet Explorer 6** では、幅広いインターネット セキュリティとプライバシー標準を導入しています。



Example for Transfer Information (cont)

- Internet Explorer 6 では、幅広いインターネット セキュリティとプライバシー標準を導入しています。

M **A private standard has been introduced as the Internet security whose it is wide range in InternetExplorer6.**

○ **A private standard and a wide range internet security has been introduced in InternetExplorer6.**

A private standard <TAR> has been introduced as the Internet security <TAR> whose it is wide range <MOD: Internet security > in InternetExplorer6 <EXP>.

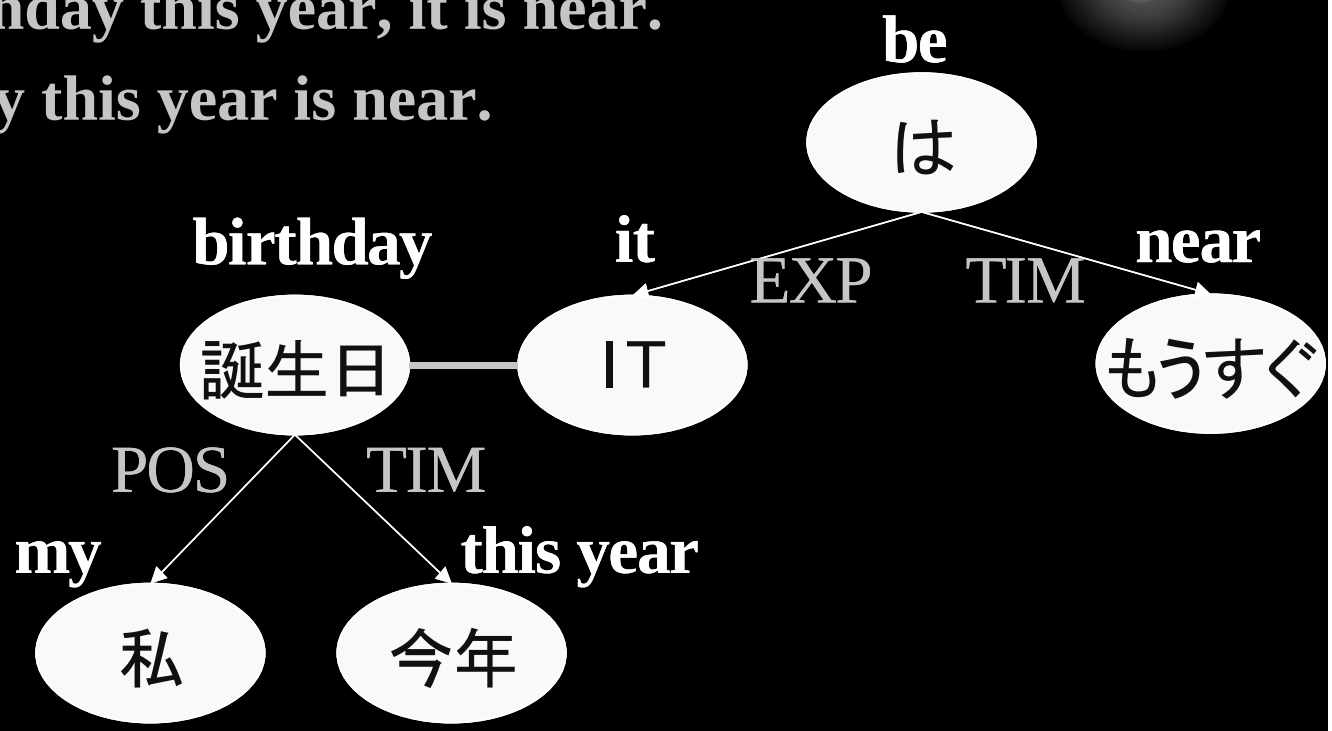
Example for Transfer Information (cont)

Additional of Words and Phrases

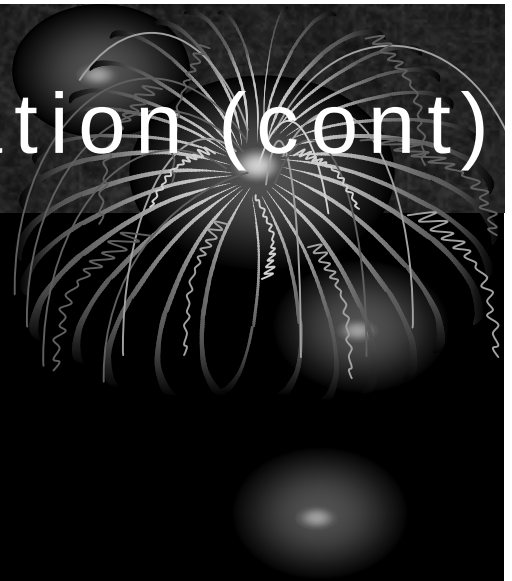
- **Ex3)私の今年の誕生日はもうすぐだ。**

  M　On my birthday this year, it is near.

  ○　My birthday this year is near.



Example for Transfer Information (cont)



- **Ex3)私の今年の誕生日はもうすぐだ。**
  M　On my birthday this year, it is near.
  ○　My birthday this year is near.

“ It ” represents “On my birthday this year”

On my<POS> birthday<TAR> this year<TIM>,
it<node: birthday> is near<TIM>.

Example for Transfer Information (cont)

Omission of words

Hamma de booru wo wattekudasai.

- Ex4) ハンマーでボール割ってください。

M　Break a ball by the hammer

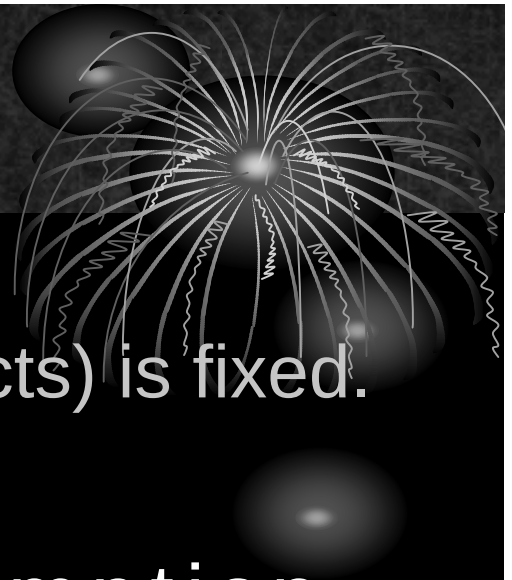
○　Please Divide a ball by the hammer

- Polite form is defined as an attribute of verb is not used
- A word “Please” is omitted

Break <c#divide;style:polite> a ball <OBJ>

by the hammer <INS> .

Conclusions: Benefits



- Reducing Digital Divide
  - The gap among languages (or dialects) is fixed.
  - More information is shared.
- Reducing Cost, Time Consumption
  - Local language←→ English Bilingual System
  - Possible for all Types of MT System

    Meteo : French ←→ English MT System

    (similar languages)

    Parsit : Thai ←→ English MT System

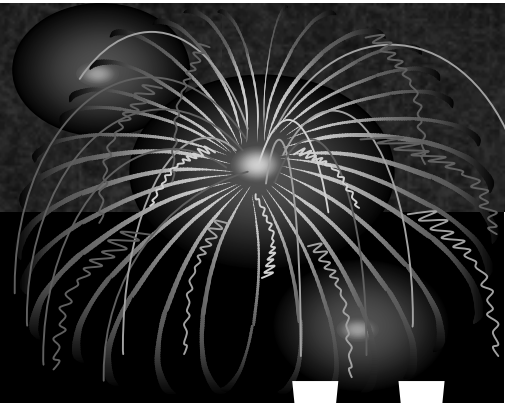
    (non-similar languages)
- Applying to Various Application Tools

Conclusion: Still problem!!

- Appropriate transfer information
  - The more knowledge we transfer, the more accuracy will increase. (??)
  - In which problem, which knowledge is appropriate to transfer!!
- Approach level
  - How well of cross system MT should be?

Future Work

- Investigate essential knowledge for each problem
- Apply this idea to the existence MT system
  - For example) T→ E → J and J → E → T
    - We will get T→J and J→T
- Apply to some application tools
  - Information retrieval



Thank you!!

Contact Us:
Thepchai Supnithi
Thatsanee Chareonporn
Virach Sornlertlamvanich
National Electronics and Computer TEchnology Center
Information Research and Development Division
thepchai@nectec.or.th

The 20 Most Spoken Languages in the World

Position	Language	Family	Script	Speakers(million)
1	Mandarin	Sino-Tibetan	Chinese	900
2	English	Indo-European	Latin	430
3	Hindi	Indo-European	Devanagari	320
4	Spanish	Indo-European	Latin	310
5	Russian	Indo-European	Cyrillic	280
6	Arabic	Afro-Asiatic	Arabic	185
7	Bengali	Indo-European	Bengali	180
8	Portuguese	Indo-European	Latin	175
9	Malay/Indonesian	Malayo-Polenesian	Latin	140
10	Japanese	Altaic	Chinese/Japanese	125
11	German	Indo-European	Latin	120
12	French	Indo-European	Latin	115
13	Urdu	Indo-European	Nastaliq	88
14	Punjabi	Indo-European	Gurumukha	75
15	Korean	Altaic	Hangul	68
16	Telugu	Dravidian	Telugu	64
17	Italian	Indo-European	Latin	63
18	Tamil	Dravidian	Tamil	62
19	Matathi	Indo-European	Devanagari	61
20	Cantonese	Sino-Tibetan	Chinese	60