Today's topics

- Japanese written language
  - Isolating factor
  - Complicates IT processing
- Terminology issues (homonyms)
- Language structure
  - Adds challenges to MT (as if more were needed)
- Impact on translation industry in Japan

Kanji characters

🐟 → 🈁 → 🈁 → 魚

鱈

鮭

鰊

Character for "fish":

Sakana, uo, gyo

Uo hen: left side fish radical

Kanji compunds

- 漁業　gyogyou fishing industry
- 魚類額　gyoruigaku ichthyology
- 魚串　uogushi   fish skewer
- 魚屋 sakanaya fish store
- 焼き魚　yakizakana grilled fish
- 熱帯魚　nettaigyo tropical fish
- 川魚 kawauo river fish
- 雑魚 zako small fish, small fry

Inflections of verbs and adjectives

- 食　　character for "eat"
- 食べる taberu eating
- 食べました tabemashita ate
- 食べませんでした tabemasen deshita
didn't eat
- 食べられました taberaremasita   was eaten
- 食べなさい tabenasai eat!

Hiragana phonetic characters evolved from kanji used for inflections

あ   a   阿、安、英、足

い   i   伊、怡、以、異、已、移、射、五

う   u   宇、羽、于、有、卯、烏、得

え   e   衣、依、愛、榎

お   o   意、憶、於、應

Input technologies

- Analog
  - Hand picking/imaging char. by char. on grid of thousands of characters
- Keyboard input with hiragana (50 keys or roman input) then conversion to kanji
  - Choosing among homonyms
  - Word parsing challenges
- Cell phone input (approx. 15 keys)
  - New predictive input engines

Visualizer

- Phototypesetter (1/1 images)

Character sets

- Japanese Industrial Standards
  - JIS X 0201 (Roman characters and katakana)
  - JIS X 0208 (basic Japanese character set)
  - JIS X 0212 (supplemental character set)
  - JIS X 0213 (recent extension to JIS X 0208)
- Universal character set (Unicode)
  - Contains characters for all key writing systems of the world
  - Includes all characters from JIS X 0208, JIS X 0212, and JIS X 0213

Character encodings

- Japanese character encodings
  - Shift-JIS
  - EUC
  - JIS
- Unicode character encodings
  - UTF-8, UTF-16, UTF-32
- Transcoding

Homonyms

Word 1	Word 2	Description
しょう	勦	destroy
しょう	剿	destroy
しょう	倡	actor
しょう	上	upper part/government
しょう	賞	prize/award
しょう	象	phenomenon
しょう	章	section/medal
しょう	称	call/label (vs)
しょう	症	illness (an)
しょう	召	call/send for/wear/put on/take (a bath)/ride in/buy/eat/drink/catch (a cold)
しょう	匠	workman/artisan/means/idea
しょう	償	make up for
しょう	傷	wound/injury

Kana to kanji conversion example

いよいよさいごのそうしゃです

⬇
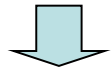
いよいよ最後の走者です

OR

いよいよ最後の奏者です

No word delimiters

- No spaces between words
- Hiragana symbols are used to inflect the ends of verbs and adjectives
- Pattern of kanji and hiragana assists the human reader in separating distinct words
- Computer applications segment text into word units through dictionary lookup, morphological analysis, and statistical techniques

Parsing challenges

- きょう**はは**をたずねた
- 今日母を訪ねた

- きょう**はは**をけずられた
- 今日は歯を削られた

Line wrapping rules

- No spaces between words
- Lines can break in the middle of words
- *kinsoku shori* rules govern line wrapping
  - Characters which cannot begin a line
    、，’”）〕］｝〉》」』】・：；？！。．　°′″％‰￠ヽバゝゞ々ー
  - Characters which cannot end a line
    ‘“（〔［｛〈《「『【￥＄£

Sorting

- Kana are ordered by table of 50 sounds
- Kanji can be ordered by representative reading, radical, and stroke count
- Culturally appropriate collation algorithms use multiple levels
- In software applications a phonetic field is often associated with Japanese names so they can be sorted by pronunciation

Searching

- Lack of word delimiters
- Redundant writing system
  - Hiragana and katakana syllabaries
  - Half-width and full-width characters
  - Arabic and Chinese numbers
- Inconsistent orthography
  - Inconsistent transliteration of loan words
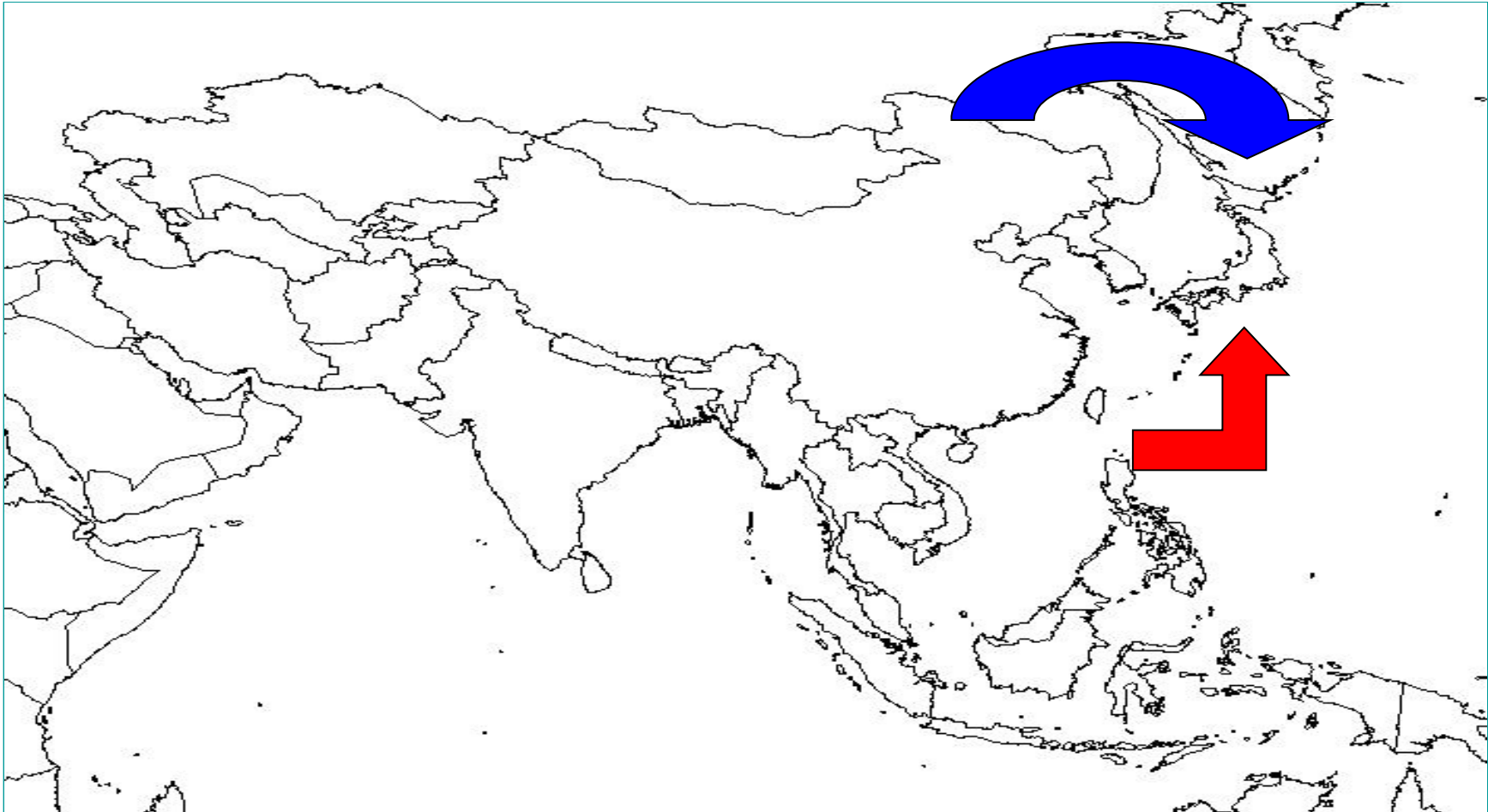 - Variation in use of kanji and kana

Furigana

- Proper nouns can have unusual and non-standard readings
- *Furigana* are small hiragana or katakana characters written above kanji to indicate the pronunciation

かんじ
漢字

LINGUISTIC DISTANCE

SOV structure

- Basic sentence pattern is: SUBJECT (often omitted)/OBJECT/VERB

- Kare ga   sakana o   tabemashita.

  He ⟵ topic marker   fish ⟵ object marker   ate

Modifying clause precedes noun

- He took the book I was waiting for.

- Kare wa   watakusi ga   matteita hon o   torimasita.
  He   topic marker   I   subject marker waited   book topic marker took.

No equivalent in this sentence of English word "for."

(Japanese learners of English and J-E MT systems often have trouble with such prepositions.)
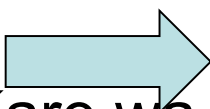
MT Example

- Kare wa watakusi ga matteita hon o torimasita.

(He took the book that I was waiting for)

"He took the book which I waited." (same result if "that" is omitted.)

He took the book that I was waiting for.

" 彼は本を取った私が待っていた ."

Kare wa hon o totta watakusi ga matte ita.

Software fails to recognize "I was waiting for" as a modifier of "book" and/or fails to place it before "book" in Japanese.

("He took the book" "I was waiting")("He[I who took the book was waiting].) Nonsense Japanese.

Strangely, software translates correctly if "that" is omitted.

MT Example (cont.)

- Same tool, Eng. to Fr. ➡ Il a pris le livre que j'attendais (same result whether or not "that" is omitted)

- Another free E-J translation web-based tool rendered the sentences fine (both E-J and J-E)

- So these easy problems are solvable but with greater effort

- Implies lower upper limit of accuracy?

Issues in help linking, TM segment alignment etc.

- Hard to force translators to obey one-to-one sentence alignment due to tendency towards awkward results
  - Sentences with abstract noun subjects often very strange in Japanese; require complete rearranging of sentence or across sentences
- Reversed position of clauses modifying nouns can complicate links

CULTURAL DISTANCE

- Interpreters who greeted Perry's ship ca. 1860

- Japan was mostly isolated 1660-1860 while the West saw massive cross-border exchange and conflict

Lack of scale in trans. industry 1

- Lack of software developed in Japan for global markets (holds down demand for high volume fast turnaround multi-language projects that force translation companies to scale up and go global)
- Lack of status of service industries in Japan (hard to get financing to grow businesses that lack tangible assets such as factories)
- Lack of qualified Japanese staff at locations around the world

Lack of scale in trans. industry 2

- Emphasis in Japan on English translation, with very low skill levels in languages such as French, German and Spanish
- Lack of multilingual content management strategy at major Japanese firms

Lack of scale in trans. industry 3

- China
  - Source of cheap translation labor?
  - First step for Japanese translation companies seeking true global footprint?

Carl Kay
Consultant
Osaka, Japan
carlkay@mub.biglobe.ne.jp

David Fine
Internationalization Consultant
Philadelphia, United States
david.fine2@verizon.net

Thank you.

ご清聴有難うございました。