Linguistics D: Day 4 Extension: Tokenizer

1 Tokenizer for English

1. English word counter is available to count and classify words in a text.
2. TreeTagger - a part-of-speech tagger for many languages

2 Tokenizer for Modern Japanese

1. MeCab: Yet Another Part-of-Speech and Morphological Analyzer
2. JUMAN (a User-Extensible Morphological Analyzer for Japanese)

3 Text generation with Markov Chain/Bi-gram

"Tom Sawyer Abroad (1894)" by Mark Twain.

The Bi-gram is a pattern consisting of the two words connected. In other words, the collection of patterns which are made of two adjacent words. The following plot is of "Tom Sawyer Abroad (1894)" by Mark Twain. The era of the Internet was ever once called the Usenet, a computer program called **Mark V. Shaney** was very popular among hakers. We try to reproduce this processing.

Listing 1 Sentences in "Tom Sawyer Abroad" by Mark Twain are transformed into Markov Chain.

```
% cat marktwain-tomsawyerabroad.txt | ~/tag/cmd/tree-tagger-english | awk '{printf␣$1␣"␣"}' | shaney | awk '{printf␣$1␣"␣"}' | sed -e 's/␣[\,\.\:]/\.\n/g' | sed -e "s/␣+n't/n't/g" |sed -e "s/␣\'t/'t/g"|sed -e "s/␣'[sS]/'s/g" | nl
```

1. TOM SAWYER ABROAD By Mark Twain 0006 We judged the woman would go and take a
2. and said he couldn't hardly hold in.
3. and you never see the yaller sand.
4. but he wanted to catch him.
5. We see this plan was a flea up to the east end of interest.
6. because if I had a good deal harder.
7. You can't mean it ! They said I couldn't have it more.
8. Every now and done it.
9. It warn't nothing but just the same in both.
10. and smoking tobacco.
11. and only our heads out and put something on it just beginning to turn out and shouts.
12. God has made me afraid.
13. too; though when you 've made a lunge for Tom never let go all holts and fell.
14. We see that they was dates.
15. because the britches was as white as snow.

Q.1 Look at the output of the above, discuss the differences between these statements and the ordinary English sentences.

Q.2 The distribution of word frequency is L-shaped. Discuss what kind of distribution 2 grams patterns have?

4 Allocation of sentence elements

1. The flexibility of word order: The language whose word order of the sentence elements have been completely fixed does not exist.(Yoshida 2001: 211)
2. On the other hand, it is not seen so much flexibility in the order of the internal element of the word or phrase; The position of the main part in a phrase or the restraint affix in the word is usually fixed.(Yoshida 2001: 228-9;215*8)
3. The system of case: (Yoshida 2001: 215-220)
4. The omission of words:
   a. John loves Mary.
   b. *Loves Mary.
   c. *John loves.
   d. *Loves.
   e. Zhangsan xihuan Lisi ma? (Chinese)
   f. Xihuan.
   g. Vado al cinema stasera. (Italian)
   h. Voy al cine esta noche. (Spanish)
   i. Nee watashi wo aishiteru. (Japanese)
   j. Mochiron aishiteruyo.

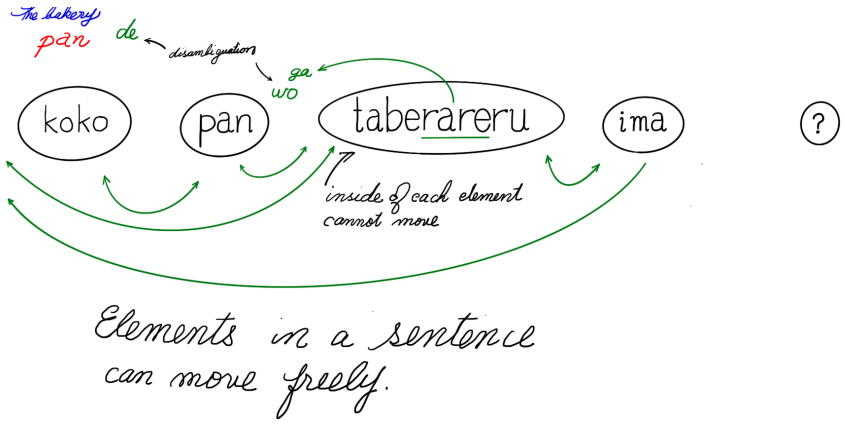

Fig. 1 Sentence and elements

Q.3 Exchange your opinion with your partner on the topic above.

5 Homework

Q.4 Access the web page from the QR code and answer the questions (deadline: today).

References

Yoshida, Tomoyuki (2001) "Nihongo wa tokushuna gengo dewa nai. (Japanese is not special language)", in Yoshifumi Hida and Takeyoshi Sato eds. *Gendai Nihongo Koza*, Vol. 1, Tokyo, Japan: Meiji Shoin, 1st edition.

Homework submission http://cuckoo.js.ila.titech.ac.jp/~yamagen/ling/linggb.html