Estimating headedness in Indonesian, French, and Finnish

Name: David Moeljadi
Matric number: G1303363A
Course number: HG2051
Course name: Language and The Computer
Lecturer: Assoc.Prof. Francis Bond
Due date: March 17, 2014, 17:00

2014
1. Introduction

In languages, the relation between elements in a phrase is either equal or dependent, i.e. one element is modified by the other elements (Tsunoda 2009: 27). In a dependent relation, one element or word can be regarded as the main, which has the same referent as the whole phrase and determines the part of speech of the whole phrase. This main element is called the "head" of the phrase and the other elements are "dependents" (Payne 2008: 31, 33, 86). For example in English, in a noun phrase "that big red house", the head is "house" and the dependents are "that", "big", and "red". Some languages, like Indonesian, tend to have the "head" in the beginning of the phrase (head-initial). For example, rumah "house" in rumah besar "big house" and kecil "small" in kecil sekali "very small". Other languages, like Japanese, tend to have the "head" in the end of the phrase (head-final). For example, ie "house" in ookii ie "big house" and chiisai "small" in totemo chiisai "very small".

The degree of headedness varies according to languages. Indonesian, although have many head-initial phrases, can have the head-final. For example, buku "book" in banyak buku "many books". Japanese too, although very rare, can have head-initial phrases. For example, ryokan "inn" in Ryokan Kawamoto "Kawamoto Inn" (Tsunoda 2009: 10). It is important to know the headedness or the order of elements in phrases for describing a language, linguistic typology, and for practical purposes such as learning a language. This paper examined the percentage of headedness in Indonesian, French, and Finnish, employing lexical database data from Wordnet (Bond and Paik 2012, Bond and Foster 2013). The Python programming language (Bird, Klein and Loper 2009) was used to process the data. Indonesian, French, and Finnish were chosen because they are from different language families. Indonesian is an Austronesian language, French is an Indo-European language, and Finnish is a Finno-Ugric language (Lewis 2009). Wordnet, which is originally from Princeton Wordnet, is a large lexical database of English, suitable for processing linguistic data through computer. Various wordnets have been created for many languages, such as Indonesian (Nurril, Sapuan and
The Python programming language was employed because it has functions for processing linguistic data and thus is suitable for Natural Language Processing (NLP).

2. Headedness in Indonesian, French, and Finnish

In Indonesian, noun phrases, adjective phrases, and verb phrases are usually head-initial. For example,

1. **ibu teman saya**
   - mother friend 1sg
   - 'my friend's mother'

2. **kopi encer**
   - coffee weak
   - 'weak coffee' (Liaw 2004: 57)

3. **rajin sekali**
   - diligent very
   - 'very diligent' (Liaw 2004: 44)

4. **pergi jalan-jalan**
   - go walk-RED
   - 'go for a sightseeing'

However, they can be head-final for nouns when they are preceded by quantifiers and numerals, for adjectives when they are preceded by adverbs of degree, and for verbs.

For example,

5. **lima rumah**
   - five house
   - 'five houses' (Liaw 2004: 2)

6. **banyak buku**
   - many book
   - 'many books' (Liaw 2004: 2)

7. **agak jauh**
   - rather far
   - 'rather far' (Liaw 2004: 45)

8. **asyik bekerja**
   - absorbed in work
   - 'absorbed in working' (Liaw 2004: 47)

French noun phrases with adjectives are usually head-initial. For example,

9. **cahier-s vert-s**
   - note-pl green-pl
   - 'green notes' (Kyoto University French Class 1993: 12)

However, they are head-final when they are preceded by quantifiers, numerals, and possessive pronouns. Adjective phrases are usually head-final and some noun phrases with adjectives are
also head-final. For example,

(10) \textit{deux an-s} \hspace{1cm} (11) \textit{mes parent-s}

\begin{tabular}{ll}
\text{two} & \text{1sg.poss.pl parent-pl} \\
\text{'two years'} & \text{'my parents'} (Kyoto University French Class 1993: 16)
\end{tabular}

(12) \textit{trop petite} \hspace{1cm} (13) \textit{beau pays}

\begin{tabular}{ll}
\text{too small.f} & \text{beautiful.m country} \\
\text{'too small'} & \text{'beautiful country'}
\end{tabular}

In Finnish, most of the phrases are head-final with cases. For example,

(14) \textit{koira-n kuva} \hspace{1cm} (15) \textit{minun nemi-ni}

\begin{tabular}{ll}
\text{dog-GEN} & \text{photo.NOM} \\
\text{'dog's photo'} (Matsumura 2005: 56) & \text{'my name'} (Matsumura 2005: 57)
\end{tabular}

(16) \textit{suuri kaupunki} \hspace{1cm} (17) \textit{kahdeksan tuntia}

\begin{tabular}{ll}
\text{big.NOM} & \text{city.NOM} \\
\text{'big city'} (Matsumura 2005: 60) & \text{'eight hours'} (Matsumura 2005: 60)
\end{tabular}

Regarding the morphology of the words, Indonesian does not mark cases, gender, and plurality. French has gender and plural markers, but not cases; while Finnish has many case markers but not gender markers. Verb phrases in French and Finnish are usually head-initial.

3. Methodology

Using the Python programming language, the wordnet file which contains single word and multiple word lexemes with their part of speeches were opened, these words were extracted and classified according to the part of speech, whether they are noun, verb, or adjective. For each group of part of speech, a list of single entry words and a list of multiple words were created, then the number of the same items in the single entry word list and in the multiple entry word list were counted. Each multiple entry word was examined whether the item in the first or in the last position is the same as the one in the single entry word list. If
single entry words appear more often in the first or initial position of the multiple entry words in a language, we may conclude that it has more head-initial feature, and vice versa. The whole program is attached in the appendix.

4. Results and discussion

Employing the method mentioned above, the program was made and executed. The results can be seen in Table 1 for Indonesian, Table 2 for French, and Table 3 for Finnish.

Table 1. Headedness of nouns, verbs, and adjectives in Indonesian

<table>
<thead>
<tr>
<th>Headedness</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head-initial</td>
<td>51.69%</td>
<td>69.13%</td>
<td>57.89%</td>
<td>59.57%</td>
</tr>
<tr>
<td>Head-final</td>
<td>48.31%</td>
<td>30.87%</td>
<td>42.11%</td>
<td>40.43%</td>
</tr>
</tbody>
</table>

Table 2. Headedness of nouns, verbs, and adjectives in French

<table>
<thead>
<tr>
<th>Headedness</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head-initial</td>
<td>67.16%</td>
<td>55.95%</td>
<td>36.52%</td>
<td>53.21%</td>
</tr>
<tr>
<td>Head-final</td>
<td>32.84%</td>
<td>44.05%</td>
<td>63.48%</td>
<td>46.79%</td>
</tr>
</tbody>
</table>

Table 3. Headedness of nouns, verbs, and adjectives in Finnish

<table>
<thead>
<tr>
<th>Headedness</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head-initial</td>
<td>33.52%</td>
<td>99.14%</td>
<td>5.71%</td>
<td>46.12%</td>
</tr>
<tr>
<td>Head-final</td>
<td>66.48%</td>
<td>0.86%</td>
<td>94.29%</td>
<td>53.88%</td>
</tr>
</tbody>
</table>

Indonesian noun phrases, verb phrases, and adjective phrases are more head-initial. French noun phrases and verb phrases are more head-initial but the adjective phrases are more head-final. Finnish noun phrases and adjective phrases are more head-final while the verb phrases are head-initial (99.14%). Overall, Indonesian has the most number of head-initial phrases (59.57%) followed by French (53.21%) and Finnish (46.12%).
5. Conclusion

In this paper, the percentage of headedness of noun, verb, and adjective phrases in Indonesian, French, and Finnish were calculated, employing the wordnet data and the Python programming language. The result was Indonesian and French have more head-initial phrases while Finnish has more head-final phrases.

References


Appendix

The program

## Program for estimating headedness of language(s) from Wordnet
## HG2051: Language and The Computer, lecturer: Assoc.Prof.Francis Bond
## student: David Moeljadi, due date: Mar 17, 2014 17:00

## program starts here
## codecs allows us to read unicode files
import codecs

## open the Wordnet file
def openFile(language):
    try:
        wnfile = "wn-data-" + language + "\.tab"
        print "\nProcessing data from: " + wnfile
        ## open each wordnet file as utf-8 text, read-only 'r'
        f = codecs.open(wnfile, encoding='utf-8', mode='r')
        ## make empty lists for each part of speech
        nounList = []
        verbList = []
        adjList = []
        for line in f.readlines():
            if line.startswith('#'): ## ignore comments
                continue
            ## strip off end-of-line, then split
            items = line.strip().split('¥t')
            ## just print the lemmas
            if items[1].endswith('lemma'):
                if items[0].endswith('n'):
                    nounList.append(items[2])
                elif items[0].endswith('v'):
                    verbList.append(items[2])
                elif items[0].endswith('a'):
                    adjList.append(items[2])
        return nounList, verbList, adjList
in case the input is not correct, go back to the start of
the program
except IOError:
    print "No such language. Please input the name correctly."
program()

make lists of words and count the headedness
def count(onePOSList):
    # create a set of single-entry words
    singleList = set(word for word in onePOSList if len(word.split(' ')) == 1)
    # create a list of multiple-entry words
    multiList = [words.split(' ') for words in onePOSList if len(words.split(' ')) > 1]
    # count the number of the same items in single-entry words and multiple-entry words
    front = 0
    back = 0
    for words in multiList:
        if words[0] in singleList:
            front += 1
        if words[-1] in singleList:
            back += 1
    frontPercent = 100.0 * front / (front + back)
    backPercent = 100.0 * back / (front + back)
    return frontPercent, backPercent

start! welcome greetings
print "*** Welcome to linguistic 'Headedness' program ***¥n" print "##Make sure the Wordnet .tab file(s) is in the folder##¥n" print "*Language choices: ind for Indonesian, fra for French, fin for Finnish*"

important! the function program
def program():
    body()
    tail()

def body():
language = raw_input("Choose a language (ind/fra-fin): ")
## open the Wordnet file and make a list for each part of speech
noun, verb, adj = openFile(language.lower())
## count the percentage of headedness
nFront, nBack = count(noun)
vFront, vBack = count(verb)
aFront, aBack = count(adj)
## count the average of front (head-initial) and back (head-final) in a language
avgFront = (nFront + vFront + aFront) / 3
avgBack = (nBack + vBack + aBack) / 3
## print the result
print "%n*** Percentage of headedness in %s ***" % language
print "%12s %6s %7s %6s %7s" % ("Headedness", "noun", "verb", "adj", "all")
print "%-12s %6.2f%% %6.2f%% %6.2f%% %6.2f%%" % ("Head-initial", nFront, vFront, aFront, avgFront)
print "%-12s %6.2f%% %6.2f%% %6.2f%% %6.2f%%" % ("Head-final", nBack, vBack, aBack, avgBack)

def tail():
    ## offer the option to choose another language
    choice = raw_input("%nChoose another language (y/n): ")
    ## if yes, back to the start of the program
    if choice.lower() == "y":
        program()
    ## if no, program ends
    elif choice.lower() == "n":
        print "%nEND OF PROGRAM"
    ## in case the input is not correct, go back to "choose another language"
    else:
        print "Please input correctly."
tail()

## execute the program!
program()