Project 1, HG7030
March 2014

Using wordnets to investigate headedness

Bruno Olsson

1 Background

1.1 Introduction

This paper reports on an investigation using wordnets for 4 languages in order to study the position of the head word in the NP, AdjP and VP. The languages belong to different families, as shown in Table 1.

Table 1: Languages of the wordnets.

code	name	family	reference
BAS	Basque	isolate	Gonzalez-Agirre et al. 2012
ENG	English	Germanic	Fellbaum 1998
FIN	Finnish	Finno-Ugric	Lindén and Carlson 2010
IND	Indonesian	Malayo-Polynesian	Noor et al. 2011

1.2 Wordnet

The study employs wordnets that are freely available for research (see the references in Table 1; the wordnets are made availabe in a uniform format by Bond and Paik 2012). The entries in the wordnets were sorted according to their POS-tags (noun, adjective or verb), and then a second time according to whether they are single-word expressions (SWEs) or multi-word expressions (MWEs). Each MWE was checked to see whether the first word, final word or both of these corresponded to the same POS as the expression in question (e.g. noun for NP). If neither corresponded to a SWE tagged with the same POS, the MWE was coded as "other". This classification was used to estimate whether the phrases of the language are head initial or head final. The question of headedness is of major importance in theoretical syntax and typology, as discussed in the next section.

1.3 Phrases and their heads

The notion 'head' occupies a time-honoured place in linguistic theory but has resisted attempts at rigorous definition. In a classic paper, Zwicky (1985) decomposed the term head into eight different properties that a head can display, e.g. being the semantic argument; being the governing element (determining the morphological shape of its coconstituents); or being distributionally equivalent to the phrase as a whole. It is well known that such criteria will yield different results for different phrases (NP, PP, etc.), and various arguments have been made for the "correct" analysis that would permit the syntactic notion of head to be rescued (e.g. Hudson 1987). In the present study, however, the theoretical concerns are of less importance since the identification of heads is made entirely from the information available in the wordnets. For a multi-word expression such as *hot dog*, the program looks for single-word expressions tagged as the same POS and consisting of either of the two constituents; it finds *dog* as a single-word constituent and lists this as the head. Thus, we assume the classification made in the wordnet for the present purposes.

Despite the uncertainties surrounding heads in syntax, typological studies concerning the order between the head and its dependents have been thriving during the last five decades or so. Important contributions are the chapters by M. Dryer in *WALS* (Dryer and Haspelmath 2013), which display the order of head and dependent in a number of construction (verb and object, adjective and noun,

Table 2: Basque.

	% first	% last	% both	% other	*n*
NP	24.2	31.1	34.0	10.7	3201
AdjP	0.0	25.0	0.0	75.0	4
VP	0.2	97.1	1.2	1.6	2633

numeral and noun etc.) in a large language sample. However, Dryer's typology is largely bimodal (an example of "data-reduction typology", Wälchli 2009) in that it reduces the complex situation that is found in actual languages to two values (e.g. verb before object vs. object before verb). Language typically do not obey such strict classification. This can be seen in the examples from Indonesian below, where AdjPs are found with degree modifiers both before and after the head (1), or NPs with a noun modifier before or after the head (2). (The heads, marked in bold below, are identified on a purely semantic basis in these examples).

(1)	sangat **buruk**	'extremely bad'
	jahat sekali	'very mean'
(2)	**fluks** neutron	'neutron flux'
	ibu **kota**	'capital city' (lit. *mother city*)

In Dryer's approach, a decision has to be made as to which of the patterns is the most 'basic', according to productivity, frequency and other criteria, and the language is coded accordingly.

A different approach would be to regard the variation in order as an interesting typological fact in itself. This is what is attempted here.

2 Results and discussion

2.1 Basque

The results derived from the wordnet for BAS are shown in Table 2. The percentages in the two first columns show how many of the first and last elements of MEWs also occured as SWE within the same POS, so that they can be assumed to be the head of the expression. The third column shows the percentage of MEWs where both the first and the last member occured as SWEs so that the head status could not be determined. The fourth column shows cases where neither the first nor the last members occured as SWEs within the same POS. The column marked *n* shows the total number of MWEs for each POS.

The results show that BAS NP-internal word other is rather heterogenous, with no clear preferencefor head first/last. This is probably because adjectives and demonstratives follow the head inside the NP, while genitives, numerals and relative clauses precede it (see datapoints for BAS, chaps. 81–91 in Dryer and Haspelmath 2013). Unfortunately, the method used here can not distinguish between the different types, so no further conclusions can be drawn.

Only four multi-word AdjPs were found in the wordnet, which is to little for any conclusions. VPs are almost excusively head-final (97.1%), which is consistent with the verb-final character of the language.

Table 3: English.

	% first	% last	% both	% other	*n*
NP	13.0	22.7	58.8	5.5	62408
AdjP	32.6	20.1	29.3	18.0	522
VP	39.0	1.4	58.9	0.6	4375

2.2 English

The results derived from the wordnet for ENG (Table 3) show that the method fails to identify the head for a majority of the NPs (58.8%). This is probably due to the large number of noun–noun expressions in the wordnet, such as *emergency landing* or *apartment house*. These are clearly head final, but the method can not identify them as such. Among the NPs that could be classified as head initial/final, it is interesting that a relatively large part turned out as head initial. Informal inspection of the data suggests that some of these are lexicalized deverbal expressions such as *face lifting*, which are incorrectly classified as head first. This also happens to lexicalized plural NPs such as *natural resources*, since the program does not recognize plural forms at present.

Again the number of multi-word AdjPs is small ($n = 522$). The results show that the order inside the AdjP varies much. For example, if there is a prepostional complement, it will follow the adjective, as in *contrary to fact* (note however that their distribution is restricted, and they can not modify a noun: **a contrary to fact statement*). Adverbial modifiers, however, precede the adjective, as in *politically incorrect*. As expected, many MWEs have adjectives both in the end and the beginning (e.g. *increasing monotonic*) and can not be correctly classified by the program.

A large part of the multi-word VPs seem to be head-initial (39.0%), which is expected given the large number of lexicalized verb + preposition/particle combinations in ENG: *carry on*, *gloss over*, *get along* etc. A problem is that several prepositions seem to be listed as verbs, so that MEWs with e.g. *out* and *up* (such as *turn out* and *loosen up*) are classified as having verbs in both initial and final position. This is a problem with the classification made in the wordnet, and shows up in the large percentage of verbs classified under "both" (58.9%).

2.3 Finnish

The NP in FIN is mostly head final, for example *akateeminen lukuvuosi* 'academic year' or *juomakelpoinen vesi* 'potable water'. Inspection of the data suggests that many of the MWEs labelled as "other" result from the program's inability to recognize case-inflected nouns. One example is the NP *kertaa minuutissa* 'times per minute' in which the head *kertaa* has the Partitive case ending -*a* required after (plural) numerals, and the dependent *minuutissa* has the Innessive case ending -*ssa*. Such forms can not be identified at present. The same problem also affects the classification of AdjPs.

Multi-word VPs are overwhelmingly verb initial, even when the only argument is a semantic subject, e.g. *sataa lunta* 'to snow' (lit. 'fall snow').

2.4 Indonesian

Indonesian (Table 5) resembles English in that many multi-word NPs both start and end with a noun, as the examples in (2) above, resulting in a high score (49.2%) for this category. Interestingly,

Table 4: Finnish.

	% first	% last	% both	% other	*n*
NP	17.5	47.1	12.7	22.8	31085
AdjP	2.6	69.7	1.7	25.9	2891
VP	97.8	0.1	0.8	1.3	6512

Table 5: Indonesian.

	% first	% last	% both	% other	*n*
NP	22.6	17.9	49.2	10.2	27414
AdjP	36.5	18.6	29.1	15.8	3536
VP	53.0	5.6	32.6	8.7	4460

the percentage of NPs classified as noun initial vs. noun final is fairly even (22.6 vs. 17.9), which makes it likely that there is a more variation than suggested by Dryer's classification of Indonesian as having the noun before genitives and adjectives (chaps. 86 and 87 in Dryer and Haspelmath 2013). The result for AdjPs as being mostly head initial (36.5%) also goes against Dryer's classification of Indonesian as having the order degree word+adjective (chap. 91 in *WALS*), again suggesting more variation.

3 Conclusion

Although the investigation yielded some interesting results—as when the scores for Indonesian showed a less clear-cut picture than the binary *WALS*-classifications—it is clear that "automatic wordnet typology" is not yet at a stage where it can offer insights not available from traditional typology. One of several issues is that wordnets are in no way representative of language use as they are lexicons and not corpora. An interesting possibility for future research would be to use the classifications arrived at here in combination with parallel texts for the different languages (e.g. translations of the New Testament) to measure the occurences of head initial vs. head final phrases in texts. This would provide a new, fine-grained way of measuring word order differences between languages.

References

Bond, Francis and Kyonghee Paik. 2012. "A survey of wordnets and their licenses". In: *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue, pp. 64–71.

Dryer, Matthew S. and Martin Haspelmath, eds. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `http://wals.info/`.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Gonzalez-Agirre, Aitor et al. 2012. "Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base". In: *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Hudson, Richard A. 1987. "Zwicky on heads". In: *Journal of Linguistics* 23.1, pp. 109–132.

Lindén, K. and L. Carlson. 2010. “FinnWordNet: WordNet på finska via översättning”. In: *LexicoNordica – Nordic Journal of Lexicography* 17, pp. 119–140.

Noor, Nurril Hirfana Mohamed et al. 2011. “Creating the open Wordnet Bahasa”. In: *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25).* Singapore, pp. 258–267.

Wälchli, Bernhard. 2009. “Data reduction typology and the bimodal distribution bias”. In: *Linguistic Typology* 13.1, pp. 77–94.

Zwicky, Arnold M. 1985. “Heads”. In: *Journal of Linguistics* 21.1, pp. 1–29.

4 Appendix: the code

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
'''This program reads in wordnets, sorts entries according to POS and checks if they
are multi-word (MWEs) or single-word expressions (SWEs). For each category, it checks
whether the first or last word, or both/none, of the MWEs also occurs as a SWEs.
The function read_wordnet uses bits of code from Francis Bond.'''

from __future__ import division
import codecs

## put names of wordnet files below
infiles = ['wn-data-ind.tab', 'wn-data-eus.tab', 'wn-data-eng.tab', 'wn-data-fin.tab']
## put name of output file below
f1 = open('output.tab', 'w')

def read_wordnet(wnfile):
    '''
    Open the wordnet file, sort all verbs into one list, nouns into another etc.
    Return one list per part of speech.
    '''
    n_list = []
    a_list = []
    v_list = []
    with codecs.open(wnfile, encoding='utf-8', mode='r') as f:
        for line in f:
            if line.startswith('#'): ## ignore comments
                continue
            ## strip off end-of-line, then split
            items = line.strip().split('\t')
            ## sort into lists according to pos-tag
            if items[1].endswith('lemma'):
                if items[0][-1] == 'n':
                    n_list.append(items[2])
                elif items[0][-1] == 'a':
                    a_list.append(items[2])
                elif items[0][-1] == 'v':
                    v_list.append(items[2])

    return n_list, a_list, v_list

def check_head(dataset):
    '''
checks for input data what proportion has word with same POS tag as entire
item first, last or in both positions, or if none.
```

```
’’’
## create set with all single-word entries, list with all multi-word entries
    single_words = set(entry for entry in dataset if len(entry.split()) == 1)
    multi_words = [entry.split() for entry in dataset if len(entry.split()) > 1]

    MWEs = len(multi_words)
## count number of times the last/forst word in MWE also occurs as SWE
    both = 0
    first = 0
    last = 0
    unknown = 0
    for item in multi_words:
        if item[-1] in single_words and item[0] in single_words:
both += 1
        elif item[-1] in single_words:
            last += 1
        elif item[0] in single_words:
            first += 1
        else:
            unknown += 1

    try:
        return  [round(first/MWEs*100, 1), round(last/MWEs*100, 1), round(both/MWEs*100, 1),
round(unknown/MWEs*100, 1),
MWEs]
    except ZeroDivisionError:
        print ’Oops!  None were found, please check the data.’
        return 0, 0, 0, 0, 0


if __name__ == ’__main__’:
    print >> f1, ’\n*** % of the XPs with X first, last, both, or in neither position (other)
    print >> f1, ’*** MEWs = total number of multi-word expressions for each category***\n’
    print >> f1, ’langXP \tfirst \tlast \tboth \tother \tMEWs’
    ## Read the wordnet file
    for infile in infiles:
        n_list, a_list, v_list = read_wordnet(infile)
        ## Check for NPs
        x, y, z, q, r = check_head(n_list)
        ## Print statistics
        print >> f1, infile[8:-4]+’NP\t%s \t%s \t%s \t%s \t%s’%(x, y, z, q, r)
        ## Check for AdjPs
        x, y, z, q, r = check_head(a_list)
        ## Print statistics
        print >> f1, infile[8:-4]+’AdjP\t%s \t%s \t%s \t%s \t%s’%(x, y, z, q, r)
```

```
## Check for VPs
x, y, z, q, r = check_head(v_list)
## Print statistics
wprint >> f1, infile[8:-4]+'VP\t%s \t%s \t%s \t%s \t%s'%(x, y, z, q, r)
```