Lund University, Dept. of Linguistics
Working Papers 49 (2001), 70–73

Stress patterns in Swedish lexicalised phrases

Per-Anders Jande
Centre for Speech Technology, KTH
jande@speech.kth.se

Abstract

This paper reports the results from a series of studies of stress patterns in Swedish lexicalised phrases (LPs). The studies conducted were carried out with the general purpose of identifying parameters that can be used to predict stress patterns in Swedish LPs. An LP lexicon with part of speech and relative stress level indicated for each word unit in the LP entries was used for this purpose. The results show a strong correlation between the part of speech and the stress level of words in LPs. The studies also show that a large part of the LPs belongs to a small group of part of speech sequences and that these sequences have very homogenous stress patterns.

1 Introduction and background

Swedish lexicalised phrases (LPs) are generally conceived of as right-heavy regarding stress (cf. e.g. Anward & Linell, 1976). This is, however, only a very general pattern. Lexicalised phrases form a very heterogeneous group, also stress-wise. The main purpose of the studies reported in this paper was to study the stress composition of LPs and attempt to identify more specific stress patterns. The means of doing this were to investigate the correlation between some different parameters and stress level for single word units in LPs and to search for structures in LPs that can be correlated with stress patterns. An LP lexicon (Lindberg, 1996; 2000) was used as a base for the work.

The loose definition of the term *lexicalised phrase* used in the studies was that an LP is a frequently occurring string of formal word units that forms a conceptual and – to some extent – syntactic unit. No attempt was made to define the concept of lexicalised phrases more closely, but the lexicon used was simply treated as the population of LPs in the Swedish language. The units included in the lexicon are chosen largely on the basis of what can be derived from the description of Swedish LPs in Anward & Linell (1976). Some typical examples of LPs are *kila stadigt* 'go steady', *göra slut* 'break up' lit. 'make end', *hålla fingrarna borta* 'keep one's hands off' lit. 'keep the fingers away' and *göra sig löjlig* 'make a fool of oneself' lit. 'make oneself foolish'.

2 Method

2.1 General issues

The choice was made to focus on two main areas of information, part of speech (PoS) information and information about the position in the LP. The information from these sources was related to stress level information in different ways. For PoS information, both the relation between part of speech and stress level for single word units and the relation between part of speech sequences and stress patterns were examined. Also, information about part of speech, PoS context and LP position was combined and the strength of the combined information for stress prediction was examined. The studies were

all quantitative to their nature and conducted with computational methods. The specific tasks of the different information extraction programs will be apparent from the presentation of the individual studies below.

2.2 Material

All data were collected from a slightly modified version of a computer-readable lexicon with Swedish lexicalised phrases as entries constructed by Janne Lindberg (Lindberg, 1996; 2000). The most important modification of the lexicon was the introduction of a special tag for reflexive pronouns. The Lindberg LP lexicon is especially good for detecting patterns in LPs, since it includes both part of speech tags and stress tags. Three levels of stress are used in the lexicon, primary stress (tagged *acc*), secondary stress (tagged *bi*) and no stress (tagged *ob*). According to Lindberg (2000), the criteria for what was counted a lexicalised phrase at the composition of the lexicon was basically criteria that can be derived from the presentation of LPs in Anward & Linell (1976), with some modifications.

3 The studies

3.1 Part of speech and stress

In this initial study, the aim was to investigate whether word units of certain parts of speech are more prone to have a certain level of stress than other in LPs. The method for investigating this was simply to collect the PoS tag and the stress tag for each word in each LP entry in the LP lexicon, combine the two types of tags from each word to a unit and count the frequency of each combination.

The study showed a strong correlation between part of speech and stress level for words in LPs. Some infrequent parts of speech always have the same level of stress when appearing in LPs, but also the more frequent parts of speech showed strong correlations with certain stress levels. For example, 89% of the verbs in the LP lexicon have secondary stress, 95% of the prepositions and 97% of the reflexive pronouns have no stress and as many as 98% of the nouns and the verb particles have primary stress. Also, 84% of the adverbs and 79% of the adjectives in the LP lexicon have primary stress.

3.2 Part of speech sequences and stress patterns

In this study, the LPs in the lexicon were grouped according to their respective PoS tag sequences. The sequences were ordered after frequency of occurrence and the possible stress patterns of each PoS tag sequence were listed (with their frequencies). Thus, in this study, the correlation between PoS sequences and stress patterns was examined rather than the correlation between single parts of speech and stress.

Also, all stress patterns were collected together with lists of the PoS sequences that can have the stress pattern in question.

The study showed that, although the different PoS sequences of LPs are many, the large majority of LPs belong to a very small group of PoS sequences. As many as 50% of the LP entries in the lexicon belong to only eight different PoS sequences. All of these common PoS sequences also have very uniform stress patterns. In fact, *all* LPs belonging to a sequence generally have the same stress pattern. There are only a few exceptions from the general patterns and all PoS sequences do not even have exceptions. The most common type of LP is the particle verb, consisting of a verb and a verb particle. In particle verbs, the verb always has secondary stress and the particle always has primary stress.

The most common stress patterns were also shared by a large number of the LPs. The nine most frequent stress patterns were shared by 85% of the LPs.

3.3 Position in the LP and stress

This study examined the correlation between the position in the LP and the stress level. The LPs were divided into groups according to their length (number of included word units) and the stress distribution at each position in each group was determined.

The study showed that the most frequent stress level in LP-initial position is secondary stress. Since the lexicalised verb phrase (LVP) is the most common type of LP and a verb in most cases have secondary stress and always stands first in an LVP, this is expected. Secondary stress seldom occurs in other than LP-initial position. Primary stress is the prevailing stress level in LP-final position, although the words in final position are also often unstressed, at least in LPs with more than two words. In the LP-medial positions, the stress level is either *acc* or *ob* in approximately equal distribution (with a bias shifting from *ob* to *acc* the further to the right in the LP we get) and there are only some occasional words with secondary stress (*bi*).

4 Stress prediction

The results from the studies could be used to predict the stress of word units in LPs. The general pattern is that content words and particles have primary stress and verbs have secondary stress. Functional words and reflexive pronouns have no stress. However, with some simple rules for predicting when the stress level of a word is not that commonly associated with the PoS of the word, the accuracy of the stress predictions could be increased. The rules include information about the position in the LP and the PoS context. Generally, the variation consists of words belonging to parts of speech that in most cases have primary stress sometimes having secondary stress and vice versa. Parts of speech associated with the stress level *ob* to a greater extent have this stress level irrespective of position and PoS context.

Verbs constitute the most frequently occurring part of speech in the LP lexicon. As many as 32% of the words in the lexicon are verbs. The verb phrase is also by far the most common type of LP; 79% of the LP lexicon entries are lexicalised verb phrases. However, in a token check in a large newspaper corpus, the lexicalised verb phrases did not have the same high relative frequency. Instead, multi-word subjunctions and lexicalised prepositional phrases were the most common LPs token-wise.

The stress predictions for verbs in LVPs were the predictions that could be improved the most with the help of a simple rule system. In this rule system, all non-LP-initial verbs were predicted to have primary stress. The LP-initial verbs were predicted to have secondary stress if the LP in question contained a word normally associated with primary stress (i.e., nouns, adjectives, adverbs, personal and indefinite pronouns, cardinal and ordinal numbers, interjections, proper names and verb particles) or another (i.e., non-initial) verb and to have primary stress otherwise. These predictions build on the assumptions that every LP wants at least one primary stress (99.7% of the LPs in the lexicon do) and that the verb will take this stress level only if there is no word belonging to a PoS that normally takes primary stress in the LP (a primary stress default PoS). The prediction procedure can thus be summarised as below.

Is the verb LP-initial?
- No → predict *acc*
- Yes: Does the LP contain a primary stress default PoS or another verb?
  - No → predict *acc*
  - Yes → predict *bi*

When these predictions were evaluated (i.e., checked against the tags in the LP lexicon), a prediction accuracy of 98.6% was achieved. There were thus clear rules for the stress level of verbs in all positions. It was also clear for parts of speech with primary stress as their "default" stress level that they have primary stress when they are the rightmost word of their kind. If they are not the rightmost word of their kind they have about an equal chance of having primary and secondary stress, respectively.

5 Implementation

Lexicalised phrase-internal structures can be interesting per se, but in general the structures in LPs are probably most interesting in relation to the "normal" rule-generated structures of the language. Comparisons to rule-generated patterns fall outside the scope of this paper in most areas and have only been made to a very limited extent. However, parts of the information gained in the studies were used to create an LP lexicon containing only LPs with "deviant" stress patterns. This lexicon is currently used for increasing the accuracy of stress assignment of a text-to-speech system in a project at the Centre for Speech Technology and Telia Promotor.

Acknowledgements

The research reported in this paper was carried out at the Centre for Speech Technology, a competence centre at the Royal Institute of Technology. I would like to thank my supervisor Beáta Megyesi for her support and comments. I would also like to thank Janne Lindberg for generously giving me access to his LP lexicon, without which the studies reported could not have been carried out.

References

Anward, Jan and Per Linell. 1976. "Om lexikaliserade fraser i svenskan" 'On Lexicalised Phrases in Swedish'. In *Nusvenska studier* 55-56:77-119

Lindberg, Janne. 1996. "Detektering av lexikaliserade fraser för text-till-talkonvertering" 'Detection of Lexicalised Phrases for Text-to-Speech Conversion'. In *the Nordic Languages and Modern Linguistics*, 191-203. Proceedings from the Ninth International Conference of Nordic and General Linguistics, Novus, Oslo

Lindberg, Janne. 2000. "Automatic Detection of Lexicalised Phrases in Swedish". In *Nodalida '99*, 103-114. Proceedings from the 12th "Nordiske datalingvistikkdager", Department of Linguistics, NTNU, Trondheim