What is Invariant and What is Optional in the Realization of a FOCUSED Word? A Cross-Dialectal Study of Swedish Sentences with Moving Focus

Robert Eklund

Telia Research AB, Spoken Language Processing
S–136 80 Haninge, Sweden
Robert.H.Eklund@telia.se

ABSTRACT

State-of-the-art speech recognition systems handle continuous speech and are speaker-independent. However, the linguistic information conveyed in the intonational contour is neglected. To be able to fully recognize speech, this information must be interpreted. To this end, explicit knowledge of dialectal and individual variation is required. In this paper some acoustic correlates of *wh*-focus in three Swedish dialects are described. Variation within and between dialects is accounted for, as well as individual differences and optional phenomena.

1. INTRODUCTION

Modern state-of-the-art automatic speech recognition (ASR) and speech-to-speech translation (SST) systems are speaker-independent and handle continuous speech with a high degree of stability. However, current systems do not make use of prosodic information. Utterances often have one or more constituents semantically focused by prosodic means, and detection of the focus in the intonational contour of an utterance and the information it conveys is important for any ASR/SST system to be able to achieve a full interpretation of the said utterance. In order to enable ASR/SST systems to identify and correctly interpret the focus/foci in a speech signal, it is vital to link a semantic model of focus to a model describing the acoustic-phonetic correlates thereof. Therefore, an understanding of invariant and optional features of focus is crucial for successful interpretation of human utterances.

1.1. A Plethora of Foci...

The term *focus* is used to denote a wide variety of different phenomena, depending on where and by whom the term is used. Not even within fairly delineated areas of research is the term used in a consistent way. In order to briefly outline the field, I will here make a (somewhat arbitrary) distinction between the meaning, or content, side of focus on the one hand, and the form, or expression, side of focus on the other.

Meaning/Content of Focus. One way to look at focus is to study its functions. What do we mean when we say that something is "in focus"? What information is conveyed by means of focus? In a somewhat simplified way, one could say that the disciplines that study these facets of the phenomenon are psychology, discourse theory, logic and semantics. Key concepts here are *information status*, *given* versus *new* information, *reference binding/resolution* and the like. One might assume that there be a measure of consistency in the use of the term *focus*, but while psychologists/cognitive semanticists tend to regard *given* information as being "in focus" (or "activated"), discourse specialists most often use the term "focus" to denote *new* information. (For a more detailed description of the different uses of "focus", cf. Gundel [6]).

Form/Expression of Focus. One may also study focus from the other side, as it were, by looking at the ways in which focus is linguistically expressed. This may be done by morphologic, syntactic or prosodic means (cf. Vallduví & Engdahl [11]). (In text, focus is most often signalled through *italics*, **bold** or CAPITAL letters.) In many languages highlighted information is often signalled by "prosodic prominence", i.e. by stressing the relevant syllable in the constituent. Terms commonly used to denote these prosodic gestures are *sentence accent*, *pitch accent*, *focal accent* and so forth. The most salient cue for stress in a number of languages has been shown to be pitch (cf. Terken and Nooteboom [10]).

1.2. ... But How Are They Related?

Clearly, there are different kinds of focus from a semantic point of view, and equally clearly, variation occurs on the acoustic side. This would imply that there is a many-to-many relation between meaning and expression foci. The interesting question this poses is whether any consistent mappings between meaning and form can be found. There have been attempts at unified approaches (e.g. Youd [12], Hoepelman [7]), and from an ASR/SST point of view, a unified model no doubt is required. In this paper, one specific kind of focus is investigated, *wh*-focus.

1.3. *Wh*-focus

Like other foci, *wh*-focus has been the object of several studies, and there is an on-going debate as to its detailed characteristics (cf. Erteshik [5]). In this paper *wh*-focus is understood in its "simplest" form, where a question word emphasizes a constituent in the reply. An example would be the question *WHO won an eel?* and the corresponding reply *A young MAN won an eel.*

1.4. Dialectal Variation

Even if we consider only one type of focus, like *wh*-focus, and even if we assume that it has but one major prosodic correlate, pitch, we still face the problem of dialectal variation since the dialects of a language often differ with respect to intonation (alongside vocabulary, phoneme inventory, morphology, syntax and so on). This means that we will probably have to account for intonationally different realizations of stress.

2. METHOD

In order to study how *wh*-focus is realised in different dialects, three sets of sentences were created. Swedish is a tone accent language, which means that all words carry either *acute* accent (Accent 1/A1) or *grave* accent (Accent 2/A2). The former is traditionally defined as a low in the main stress syllable, whereas the latter is signalled by a fall in the main stress syllable, in standard Swedish. (For further discussion, cf. Bruce [2], Lyberg [8] and Engstrand [4]). Thus, words of both accent types were included in the three sets of sentences. Set 1 contained one-syllable A1 words, set 2 contained two-syllable A1 words, and set 3 contained two-syllable A2 words (A2 needs a minimum of two syllables to be realized). To facilitate analysis, all sentences contained only sonorants. The sets are listed below.

Set 1: A1 / One syllable words
(*A young man won an eel*)
1 En UNG man vann en ål.
2 En ung MAN vann en ål.
3 En ung man VANN en ål.
4 En ung man vann en ÅL.

Set 2: A1 / Two syllable words
(*The younger man wins the eel*)
5 Den YNGRE mannen vinner ålen.
6 Den yngre MANNEN vinner ålen.
7 Den yngre mannen VINNER ålen.
8 Den yngre mannen vinner ÅLEN.

Set 3: A2 / Two syllable words
(*The young mother borrows needles*)
9 Den UNGA mamman lånar nålar.
10 Den unga MAMMAN lånar nålar.
11 Den unga mamman LÅNAR nålar.
12 Den unga mamman lånar NÅLAR.

As shown above, focus was moved between the four lexical constituents. The session leader prompted the subjects to read the sentences as if the sentences were natural responses to appropriate *wh*-question highlighting the capitalized constituents.

The recordings were made on location in Göteborg and Stockholm. In Skåne, recordings were made on four different locations: Helsingborg, Malmö, Trelleborg and Kristianstad. The material was recorded in quiet settings. (No anechoic chambers were available.) A Unix work station was used and the recordings were stored in digital form on disk. The subjects were all native inhabitants. An equal number of men and women were included. The subjects' age varied between 15 and 65.

It was found that several subjects had difficulties in reading the sentences in a natural way, many of them emphasizing more than one word per sentence—in some cases all words—a common problem associated with naïve subjects. Therefore, all sentences were listened to prior to analysis, and sentences judged as unnatural were omitted. The number of remaining speakers per utterance ranged from 7–30 for the Skåne dialects (with a fairly even distribution between the four locations), 4–13 for Göteborg and 13–31 for Stockholm.

The sentences were transcribed in *ESPS/Waves*. F_0 was then extracted using the *Enhanced Super Resolution* F_0 *Detector* (eSRFD) algorithm described in Bagshaw [1] (Fig. 1). To avoid differences caused by absolute pitch, F_0 was normalized to a musical scale given in semi-tones (Fig. 2). Finally, the mean F_0 for all normalized utterances was computed (Fig. 3). Durations were normalized and not specifically studied, since the main interest was in locating highs, lows and rises relative to the segments.

Each sentence was analyzed for all dialects. The Skåne locations were analyzed separately and as a group. The original F_0 contours were used to detect marked deviations from general intonation patterns.

3. RESULTS

Gårding [9] and Bruce & Gårding [3] provide intonational typologies for some Swedish dialects, including the ones covered in this paper. In their model, the main difference between the intonation contours of Swedish dialects is said to be the timing of highs and lows.

3.1. General Observations

Looking at general F_0 patterns, the results by and large confirm the typologies given in Gårding and Bruce & Gårding. The Skåne dialects are signified by early peaks in A1 words, and only one peak in A2 words. Although one has to take into consideration the relatively small number of speakers from each separate location in Skåne, a small tendency for

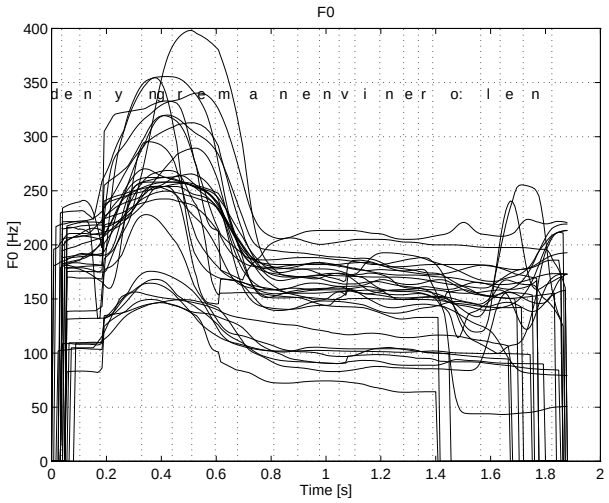

Figure 1: F_0 contours for sentence 5: *Den YNGRE mannen vinner ålen.* Stockholm speakers.

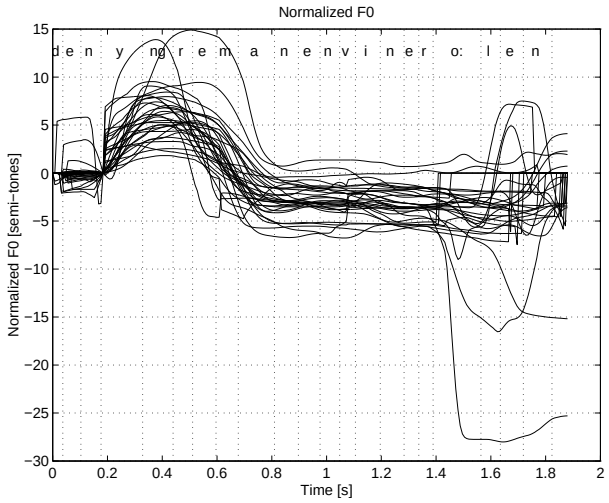

Figure 2: Normalized logarithmic F_0 contours for sentence 5: *Den YNGRE mannen vinner ålen.* The curves are fixed to a reference point at the beginning of the stressed vowel /y/. Stockholm speakers.

later peaks in Kristianstad was noted, the other three cities being well-nigh identical.

The Stockholm dialect signals focus by a high peak after the accented syllable in both A1 and A2 words. The Göteborg dialect exhibits more or less the same F_0 contours as the Stockholm dialect, showing a difference mainly in timing. Göteborg F_0 contours are realized later than the corresponding Stockholm F_0, relative to the segments.

However, since the exact location of the peak varies in the dialects under study, focus in A1 words mainly seems to correlate with a rise in the main stress syllable. In A2 words,

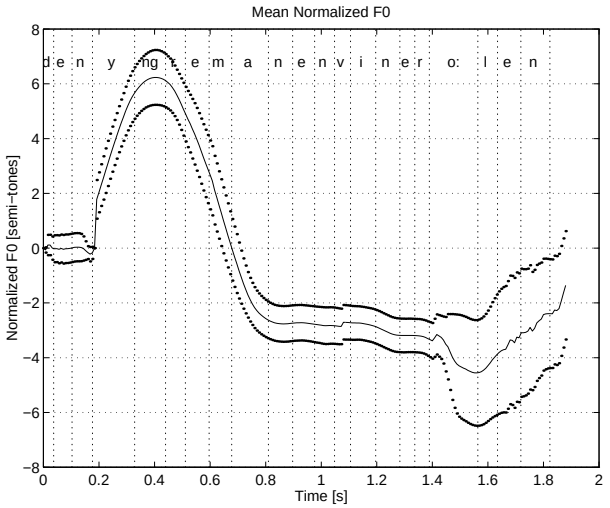

Figure 3: Mean normalized logarithmic F_0 contours for sentence 5: *Den YNGRE mannen vinner ålen.* Stockholm speakers. The outer curves denote a 95 % confidence interval.

the rise occurs in the following secondary stressed syllable following the main stressed syllable. The peak might appear in the same syllable as the rise or later, and even in another subsequent word (cf. paragraph 3.3).

3.2. Persistent Default Accent?

Although most post-focal accents were deaccentuated, there was also a small but significant tendency to retain the default accent on the last word (Figs. 2 & 3). This tendency seemingly grew with the distance between the accented word and the final word. In order to test the significance of these "two-peak" utterances, an informal listening test was conducted. A few different realizations of sentence 9, *Den unga MAMMAN lånar nålar*, were chosen, that varied between total, asymptotic post-focal deaccentuation and very clear "secondary stresses" on the last word, *nålar*. Intermediate forms were also included. The sentences were played to 15 subjects after the "cue questions" *WHO borrows needles*, (Q1) *WHO borrows WHAT* (Q2) and *Who borrows needles* (Q3), uttered by the author. The subjects were asked to rate the naturalness of the subsequent reply which was played from disk. This preliminary test suggests that people are rather sensitive to secondary peaks. Following Q1, the majority of the subjects rated sentences with secondary peaks as less natural than the corresponding sentences with clear post-focal deaccentuation. Two-peak utterances were considered more natural after Q2, where deaccentuated replies were judged as unnatural. However, these results may at best be taken as suggestive and a more controlled experiment is needed in order to draw more far-reaching conclusions on the matter.

3.3. Optional Strategies

One feature was noted where the speaker seemingly is presented with alternative options. When realizing main stress, there seems to be two "strategies" at hand. Let us use sentence 3, *En ung man VANN en ål*, as an example. While a vast majority of the Stockholm subjects execute a rise in the vowel of *VANN*, the subsequent fall is realized in different ways. Some subjects fall very quickly, already in the (phonologically long) /n/, thus exhibiting a very "narrow" peak, whereas other speakers do not execute a fall until the vowel of the next content word, *ål*. Evidently, there is no semantic difference associated with the two contours. As a matter of fact, untrained listeners cannot tell the difference between the two realizations by just listening to utterances.

3.4. Focus Glottalization

In the words with initial vowels (*ung*, *unga*, *yngre*, *ål*, *ålen*), focus was often associated with an initial glottalization. Although this was seen in all dialects (represented as vertical lines for some of the Stockholm speakers in Fig. 1) the tendency was more marked for the Skåne dialects, where very strong glottalization appeared. However, glottalization is not an "either-or" parameter, and in-between very clear cases of glottalization on the one hand and continuous voicing on the other, a group of less clear-cut cases appeared.

4. DISCUSSION

The observations reported in this paper show that focus stress is realized in different ways. Depending on the dialect—and sometimes even within the same dialect—the peak is located in different places relative to the main stress syllable. A slightly more stable correlate to focus stress seems to be the rise in the main stress syllable for A1 words, in the secondary stress syllable for A2 words. Thus, the rise seems to be the most reliable cue for sentence accent detection—at least for the dialects studied here. It is an empirical question whether this strategy could be employed in other Swedish dialects. At any rate, one may assume that a simple bottom-up strategy that tries to detect certain acoustic phenomena would meet with problems. In order to obtain successful focus detection, a top-down strategy is probably called for, where the recognizer, alongside knowledge of dialectal variation, is provided with lexicon- and discourse based hypotheses as to which words/syllables are likely takers of focus accent. An optimized approach would be to use a combination of bottom-up and top-down strategies, where the recognizer provides the syntactic-semantic module with cues for lexical lookup (needed for disambiguation between Swedish A1 and A2 words), and the syntactic-semantic module gives hints to the recognizer where to look for sentence accent peaks.

Current ASR/SST systems normally cover dialectal and individual variation, since they are trained on large numbers of persons with different dialects. However, the knowledge these systems contain is implicit, and in order to detect the intonational contour of an utterance and recognize its semantic information (whatever it may be), explicit linguistic knowledge is needed. One, among many, parts of that knowledge is to be gained from studies on dialectal and individual variation.

It must not be forgotten that a person's dialectal idiom is closely tied to his/her sense of individuality. Thus, the need to cover dialectal variation cannot be underestimated.

5. REFERENCES

1. Paul Christopher Bagshaw. *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching.* PhD thesis, University of Edinburgh, 1994.
2. Gösta Bruce. *Swedish Word Accents in Sentence Perspective.* PhD thesis, University of Lund, 1977.
3. Gösta Bruce and Eva Gårding. A Prosodic Typology for Swedish Dialects. In *Nordic Prosody*, Travaux de l'Institut de Lund. Department of Linguistics, University of Lund, 1978.
4. Olle Engstrand. F0 Correlates of Tonal Words Accents in Spontaneous Speech: Range and Systematicity of Variation. *Perilus*, Nr:1–12, 1989.
5. Nomi Erteshik-Shir. Wh-Questions and Focus. *Linguistics and Philosophy*, 9:117–149, 1986.
6. Jeanette Gundel. On Different Kinds of Focus. In Bosch & van der Sandt, editor, *Focus and Natural Language Processing*, pages 457–467, Heidelberg, 1994. IBM Deutschland.
7. Jakob Philip Hoepelman. Modellbildung der Fokusintonation im gesprochene Dialog (MAFID). Technical report, Fraunhofer Gesellschaft, Institut für Arbeitswirtschaft und Organisation, Stuttgart, 1992.
8. Bertil Lyberg. *Temporal Properties of Spoken Swedish.* PhD thesis, University of Stockholm, 1981.
9. Eva Gårding. Toward a Prosodic Typology for Swedish Dialects. In K.-H. Dahlstedt, editor, *The Nordic Languages and Modern Linguistics 2*, pages 466–474. Almqvist and Wiksell, Stockholm, 1975.
10. Jacques Terken and Sieb G. Nooteboom. Opposite effects on verification latencies for given and new information. *Language and Cognitive Processes*, 2(3/4):145–163, 1987.
11. Enric Vallduví and Elisabet Engdahl. The linguistic realisation of information packaging. *(To appear in Linguistics).*
12. Nicholas John Youd. *The production of prosodic focus and contour in dialogue.* PhD thesis, The Open University, 1992.