*Bryan Mosey**

A study on nuclei in the SEC - report from the Public Speaking Project in Lund

Abstract

As part of the Public Speaking Project in Lund, the Spoken English Corpus has been processed in a number of ways and put into a database format thus allowing the retrieval of accurate statistical data. Based on a simple definition of nucleus provided by Gerry Knowles, the nuclei in the prosodic transcriptions of the material have been located and marked. A field indicating the position of each word within the tone unit (1 representing the last word of the tone unit and 2, the penultimate word etc.) has also been added thus facilitating the statistical study of nucleus position. Also studied are the tone types occurring in the nuclei in the material. Here, the findings provide empirical support for the postulation of a dichotomy of tone types. The functions of the various nucleus types in the text categories making up the material are discussed on the basis of empirical evidence from the material.

Over the past three years, it has been a pleasure for me to be employed on the Public Speaking Project (or PS for short) at the Department of English in Lund. This project is funded by the Swedish Council for Research in the Humanities and Social Sciences and is led by Professor Jan Svartvik. The goals and background of the project are outlined in Jan Svartvik's introduction to the project (Svartvik 1991).

1. A brief description of the PS database

Before I go on to talk about my recent study of nuclei, I would like to take a few moments to describe the database we use on the project. This database consists of most of the SEC, a corpus compiled at the University of Lancaster in conjunction with the Speech Research Group at the IBM UK Scientific Centre (see Taylor & Knowles 1988). As our project is primarily interested in monologue, we have not included texts

* *Bryan Mosey*
Lund University
Department of English
Helgonabacken 14
223 62 Lund (S)

J02-J06. Category J in the SEC is labelled as dialogue. Texts J02-J05 consist of "contrived" dialogues from radio language programs while text J06 consists of an informal dialogue between two students. We found, however, that J01 was suitable for our project as it consists of a collection of short monologues with extremely little interaction between the various speakers. The text consists of a radio programme reviewing the sporting highlights of the year 1986 and is of virtually the same format as text F04 which is actually a programme in the same series and with the same "anchorman".

The database first consisted of two transcriptions of the corpus, the tagged version and the prosodic transcription. In order to be able to join these in one large database, it was first necessary to carry out a little editing. This was due in part to the fact that two analysts are responsible for the prosodic transcriptions of the material. With a number of the texts, one analyst has analysed the first part of the text and the other the latter part with an overlap in between. There are in all 24 overlap passages totalling 4680 words (see Taylor & Knowles 1988:19). To make the editing as straightforward as possible, the part which came first in the text files originally sent to us by Gerry Knowles was kept and the entire overlap extracted from the second part. This meant that for each text, we had one continuous transcription which could then be included in a database format in parallel with the other transcriptions.

The information contained in the prosodic and tagged transcriptions provided the information contained in the original eight fields of the database (fields 1-6 and 10-11 in Figure 1 below). Fields 1-4 contain locational information; fields 5 and 6, the grammatical tag (SEC employs the same CLAWS tagset as LOB); field 9, the prosodic transcription; and field 10, tone unit boundary markers. The database is organised vertically, so the information in each record corresponds to one word in the running text. It was necessary to place the tone unit boundary markers in a separate field as we wanted to mark the final record for each tone unit in order to be able to insert information relevant to a whole tone unit in the final record of that tone unit.

As work has progressed on the project additional fields have been added to include information resulting from the computerised handling of the material or from human analysis of it. Fields 12, 13, 14, 22, 23, 25 and 26 are the result of various aspects of my studies regarding prosodic features while fields 16, 17, 18, 19, 20 and 24 have come about in

connection with Olof Ekedahl's work of a more grammatical nature (see e.g. Ekedahl 1992 and Svartvik, Ekedahl & Mosey 1994). The information in a number of these has come about through Olof's parsing of the material using an adapted version of Mats Eeg-Olofsson's parser (described in Svartvik 1990 and Eeg-Olofsson 1990).

Figure 1: The structure of the database used on the Public Speaking Project[1]

[1] For reasons of space, I am unable to go into greater detail here regarding the contents of all the fields in the database but if anyone would like to know more, they are welcome to contact either Olof Ekedahl or myself (e-mail: Olof.Ekedahl@englund.lu.se or Bryan.Mosey@englund.lu.se).

2. A study of nuclei in the SEC

2.1. The definition and extraction of nucleus

In studies on prosody, a central concept is that of the tone unit, also termed intonation unit, intonation group or tone-group by various prominent linguists. Cruttenden (1986:75 ff.) provides a comprehensive guide to the features of such units. Further, he states that each such unit (he names them intonation-groups) has by definition one nucleus (Cruttenden 1986:80), the nucleus being "that pitch accent (usually the last) which generally stands out as the most prominent in each of the typical tonal sequences within intonation-groups". Closely related to Cruttenden's definition is that given by one of the compilers of the SEC, Gerry Knowles (Knowles 1993:158) who simply states that " . . . the last accent [in the tone unit] is by definition the nucleus".

Following this definition, Olof Ekedahl wrote a dBASE program which selected the nucleus from each tone unit in the material and updated the database used in this study. The program searched backwards within each minor tone unit until it found a tone marking indicating accent. Thus, stressed but not accented markings were disregarded.

2.2. Nucleus position

With regard to the definition of nucleus given above and the format of the database used in this study, the most straightforward way of quantifying nucleus position is in terms of the number of words from the end of the tone unit. Syllables have also been suggested as a suitable measure and, indeed, from the phonetician's point of view, this would probably be a more accurate measurement. However, no program was available to me that would count syllables and be able to work in conjunction with the database format. Counting in words from the end of the tone unit is also the measure used by Altenberg (1987:163) and my using the same system is useful for purposes of comparison.

Statistics on nucleus position measured in this manner will, naturally, be affected by tone unit length. For example, 619 tone units in the material consist of only one word while tone units of more than 12 words in length are rare. Graphically, the variation in tone unit length in the material offers no surprises, the curve shows the expected distribu-

tion with a peak at three words. This curve resembles that given by Altenberg in a study of LLC text S12.6 (the master builder from Stoke Poges) which is, like the texts in my material, a monologue (Altenberg 1987: 26). Average tone unit length in the material is 4.34 words with a standard deviation of 2.34 words, giving a standard deviation interval ranging from 2.00 words to 6.68 words.

To give an idea of the relationship between tone unit length and nucleus placement, the average position in which nucleus occurs can be plotted against tone unit length. The resulting curve shows a clear upward trend to about 9 words per tone unit. It then remains fairly level to about 11 words per tone unit after which it falls off. As long tone units (let us say about 12 words or longer) are rare, the data for these seem statistically unreliable. The end point of the curve, for instance, shows that tone units of 20 words have, on average, the nucleus on the second to last word. This is however based on only one example[2]:

---

2 Examples from prosodic transcriptions given in whole major tone units, relevant minor tone units in bold face and nucleus in italics. Abbreviations for intonation markings are as follows:

[L-]	Low level	[Hf]	High fall	[Lf]	Low fall
[H-]	High level	[Hfr]	High fall-rise	[Lfr]	Low fall-rise
[D]	Step down	[Hr]	High rise	[Lr]	Low rise
[U]	step up	[Hrf]	High rise-fall	[-]	Stressed but not accented
|	Tone unit boundary			||	Major tone unit boundary

F04	0012	well you're [Hf]right Kevin |
F04	0013	not only a [Hfr]jumbo-sized |
F04	0014	[Hfr]record book |
F04	**0015**	**but [H-]also a couple of [Hf]calculators as [Hf]well 'cause you [-]really needed [L]those rather than just a [H-]straightforward *[Hf]score* book |**
F04	0016	there've been [H-]so [Lf]many test [-]matches in the last 12 [-]months ||

At first glance, one might think that some mistake has been made in the transcription here, that it surely cannot be possible to cram so much into one tone unit. However, one must also take into consideration that the speaker is Chris Florence who is the fastest of all the speakers in the SEC with an average speech rate of 3.8 words per second (see Mosey 1992:11). Also, looking at this tone unit, one suspects that an LLC transcriber, for example, may have divided it into a number of so-called subordinate tone units.

The following two examples illustrate how nucleus position can vary in tone units that are still long but of a length which is somewhat more common (there are 108 cases of 11-word tone units in the material):

G01	**0586**	**a [-]hundred and *fif[Hf]teen* he [-]heard himself [-]say in[-]side his [-]head ||**
G01	**0620**	**he could [Hf]not [-]swim the [H-]few feet [L-]back to the *[Hf]rock* ||**

From the curve, we can note a strong tendency for the vast majority of nuclei to occur within the last two words of the tone unit. The nucleus can, however, occur earlier in the tone unit and the longer the tone unit, the greater the range for these rarer cases thus pushing up the average. Within the standard deviation range of tone unit length, from 2.00 words per tone unit to 6.68 words per tone unit, the trend is steadily upward, the average position of the nucleus shifting from the last word of the tone unit to more than one and a half words back from the end of the tone unit.

As can be seen from table 1 below, 77.7% of tone units have the nucleus on the last word. This explains why the scale on the vertical axis in figure 2 is so limited. Only 13.3% of tone units have nucleus on the penultimate word, 5.1% on the third from last word and 2.5% on the fourth from last word. It is interesting to note in this context that Altenberg (1987:41) found that "In TUs with simple tonicity (84% of the nuclei in his material) the most frequent position of the nucleus (87%) is on the last word" (again his study is based on LLC text 12.6). Here, it is important to point out that Altenberg (thanks to the characteristics of the prosodic transcription of the LLC) was able to distinguish between simple tonicity and compound tonicity. Although Altenberg's data is at least to some degree comparable to mine, it is a pity that his study did not include any conversational material. I have not been able to find comparable data on conversation. Nevalainen (1992) provides data on nucleus types which will be very useful but again, the positioning of these nuclei is not investigated.

From the data given in table 1 it is evident that the last word of the tone unit is the default position of the nucleus. Indeed, it would appear that tone units of all lengths can have the nucleus on the last word. The single example of a 19 word tone unit found in the material, for example, has its nucleus on the last word:

F02	**0090**	**[D]and as a [Hf]seven per cent [-]increase in the basic [Hf]pension for a [Hf]single person is [-]two pounds *[Hfr]fifty* |**
F02	0091	they'd see [Hf]virtually [U][Hf]no rise at [Hf]all |
F02	0092	in No[Lf]vember ||

At the other extreme, of course, all of the 619 one-word tone units in the material must have the nucleus on the last word, e.g.:

F02	**0030**	***[H-]up* |**
F02	0031	by only [Hfr]two per [-]cent |
F02	0032	to [L-]seven [L-]pounds a [Hf]week ||

As can be seen at the bottom of table 1, the earliest position in which any nuclei are found in the material is ten words from the end of the tone unit. In three of the four tone units where this is the case, the tone unit is longer than ten words, e.g.:

D02	0596	[H-]no [Lf]wonder [-]critics of the en[Lr]lighten-ment |
D02	0597	like [H-]Rousseau and [L-]Goethe |
D02	**0598**	**were so *[Lf]horrified* by the de[-]terminism in the [-]teaching of the ma[-]terialists |**
D02	0599	and [Hf]they [-]weren't a[Lf]lone ||

In one case, however, the tone unit is exactly ten words long and, thus, has nucleus on the first word:

G01	0659	[H-]have a [-]nice [Hr]morning she [-]asked |
G01	**0660**	***[L-]laying* her [-]hand on his [-]warm [-]brown [-]shoulder a [-]moment ||**

Table 1: Position of nucleus relative to end of tone unit (whole material)

N of words from end of TU	*N*	*%*
1	8590	77.7
2	1472	13.3
3	565	5.1
4	273	2.5
5	98	0.9
6	29	0.3
7	19	0.2
8	4	-
9	1	-
10	4	-
TOTAL	**11055**	**100**

2.3. Nucleus types and the tones which combine with them

The definition of nucleus used in the present study is a rather straightforward one which does not allow for complex and compound nuclei. Whereas this study has thus far been able to locate nuclei (according to the definition referred to) a study of complex and compound nuclei along the lines of that carried out by e.g. Altenberg (1987) would involve the reanalysis of the entire material by a trained phonetician; a job which would require greater resources than those available to me. Using the established database version of the SEC used in the Public Speaking Project, however, it is possible to look for combinations of tones occurring in combination with what have here been defined as nuclear tones. In an earlier study, I compiled statistical data on nucleus types, disregarding preceding accents (see Mosey 1993: and Svartvik, Ekedahl & Mosey 1994).

Using the Public Speaking Project's database, it was possible to collect all the prosodic markings indicating accent in each tone unit and to place them in a special field in the final record of each tone unit. The result was a symbolic representation of the prosodic contour of the tone unit. An index of the database on this field was then created allowing the various contours to be noted and, if need be, counted. As might be expected however, the degree of variation was enormous. This was due to two main reasons; one being variation in tone unit length - while a two word tone unit and an 11 word tone unit may have very similar contours around the nucleus, the difference in length between the two will mean that the overall intonation contour will not match. The second reason was the intricacy of the SEC transcription - with nine different accent markings, the number of possible combinations, even if the match-up range before the nucleus were limited, was huge.

Obviously, it would be necessary to limit the degree of variation in the analysis of prosodic contours in connection with nucleus. The first way in which this was done was by conflating the accent markings by abandoning the high/low distinction. I feel that this was a justifiable move for two reasons: 1) the distinction is not made in the LLC, and 2) the distinction between high and low in prosodic analysis is by no means a clear cut one - with regard to the SEC, data on transcriber differences in the prosodic analysis are available in an article by Pickering, Williams and Knowles (1993). It can be mentioned, for example, that of

292 falls transcribed as low falls by Gerry Knowles in the overlap sections in the SEC, 148 are transcribed by Briony Williams as high falls (Pickering, Willaims & Knowles 1993:68).

Further, as what I was primarily interested in was contours in connection with nucleus, I looked first at which patterns occurred when one accent prior to nucleus was included and then at more complex contours forming subtypes of the most common combinations.

The resulting primary conflated nucleus types and their share of the total number of tone units are given in table 2. As the prosodic marking 'stressed but not accented' is not included in the definition of nucleus, it has been ignored here. 'Level' in this table refers to level, accented tones previously expressed as 'high-level' and 'low-level'.

Table 2: The conflated nucleus types recognised in the present study and their share of the total number of tone units

Conflated nucleus type	*N*	*% of total N of tone units*
level	1534	13.8
fall	4920	44.2
rise	1915	17.2
fall-rise	2676	24.1
rise-fall	11	0.1
none	63	0.6
TOTAL	**11119**	**100**

As regards the combinations where one tone preceding the nucleus is included (again, 'stressed but not accented' is not included), the following general statement can be made: for each nucleus type, the most common combinations are with a preceding level tone or with no preceding tones at all. Combinations with a preceding fall are clearly the third most common. Combinations with other tones preceding the nucleus are relatively uncommon. To qualify this statement a little, it should be added that falling nuclei are preceded more often by a level tone than by nothing at all while with the other nucleus types, this relationship is reversed. It should also be added that rise-fall nuclei are so uncommon as to be statistically uninteresting.

Table 3 gives the ten most common combinations with the tone prior to the nucleus included (also those where nucleus is not preceded by any other tone) and their share of the total number of tone units. As can be seen from the table, the remaining combinations together account for only 5.6% of the total number of tone units.

Table 3: Combinations with tone preceding nucleus included and their share of the total number of tone units.

Conflated contour (last tone-marking = nucleus)	*N*	*% of total N of tone units*
level + fall	2279	20.5
fall	1772	15.9
fall-rise	1378	12.4
level + rise	945	8.5
level	763	6.9
fall + fall	729	6.6
rise	728	6.5
level + fall-rise	639	5.7
fall + fall-rise	623	5.6
level + level	573	5.2
all others	627	5.6
no nucleus	63	0.6
TOTAL	**11119**	**100**

As far as the level+fall contour is concerned, it has two subclasses worth mentioning, these being level+level+fall which accounts for 3.8% of all tone units in the material and fall+level+fall which accounts for a little less than 1% of all the tone units in the material. The simple fall and fall-rise contours cannot of course be subdivided. The subclasses of the level+rise contour are extremely small, the largest being level+level+rise which accounts for only 0.7% of all the tone units in the material.

In my continued studies of the nucleus, I will be able to use the PS database to look at the relationships between nuclei and their related contours and word-class tags, phrase types and prosodic finality.

References

Altenberg, B. (1987): *Prosodic patterns in spoken English*. Lund: Lund University Press.

Cruttenden, A. (1986): *Intonation*. Cambridge: Cambridge University Press.

Eeg-Olofsson, M. (1990): An automatic word class tagger and a phrase parser. Svartvik (ed.) 1990a, 107-136.

Ekedahl, O. (1992): Word and tag frequencies in the SEC. Department of English, Lund University.

Knowles, G. (1993): From text structure to prosodic structure. G. Knowles & P. Alderson (eds.) 1993, 145-166.

Knowles, G. & P. Alderson (eds.) (1993): *Working with speech*. London: Longman.

Mosey, B. (1992): Speech rate and tone unit length in the Spoken English Corpus. Department of English, Lund University.

Mosey, B. (1993): Pauses and nuclei in the Spoken English Corpus. Department of English, Lund University.

Nevalainen, T. (1992): Intonation and discourse type. *Text* 12: 397-472.

Pickering, B., B. Williams & G. Knowles (1993): An analysis of transcriber differences in the Spoken English Corpus. G. Knowles & P. Alderson (eds.) 1993, 60-85.

Svartvik, J. (ed.) (1990a): *The London-Lund corpus of spoken English: Description and research*. Lund: Lund University Press.

Svartvik, J. (1990b): Tagging and parsing on the TESS project. Svartvik (ed.)1990a, 87-106.

Svartvik, J. (1991): An introduction to the PS project: 'Public Speaking - effective communication in spoken English discourse'. Department of English, Lund University.

Svartvik, J., O. Ekedahl & B. Mosey (1994): Public speaking. *Creating and using English language corpora*. ed. by U. Fries, G. Tottie & P. Schneider, 175-187. Amsterdam: Rodopi.

Taylor, L.J. & G. Knowles (1988): *Manual of information to accompany the SEC corpus. The machine readable corpus of spoken English*. Bergen: The Norwegian Computing Centre for the Humanities.