LUND UNIVERSITY

Lexical structure and accenting in English and Swedish restricted texts

Horne, Merle; Johansson, Christer

1991

Link to publication

Citation for published version (APA):
Horne, M., & Johansson, C. (1991). *Lexical structure and accenting in English and Swedish restricted texts*. (Working Papers; Vol. 38). Department of Linguistics, Lund University.

Total number of authors:
2

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Lund University, Dept. of Linguistics
Working Papers 38 (1991), 97-114

Lexical Structure and Accenting in English and Swedish Restricted Texts

Merle Horne and Christer Johansson

The issue of describing identity of sense relations that are used in contexts of anaphora to express contextually given information is discussed. In text-to-speech applications it is important to model given information, both linguistically and computationally, since it is associated with tone accent patterns that differ from those on new information. The analysis is illustrated using restricted texts (newspaper stock market reports).

Introduction

A major problem for text-to-speech (TTS) systems aiming at generating natural prosody is modelling the factors that condition the various tone accent patterns that are used in discourse. In the case of English, for example, existing TTS systems have a very meagre inventory of tone accent types compared to the number that are actually used by speakers (see e.g. Gussenhoven 1984, Pierrehumbert 1980). However, just how one is to set about and model all the lexical, syntactic, semantic and pragmantic factors that condition the appropriate assignment of tone accents is not at all straightforward, particularly when one is dealing with unrestricted texts. A more tractable goal, however, is to limit oneself to the analysis of restricted texts, where the lexicon is considerably reduced and where it is thus possible to model a great deal of the lexical semantic information which is known to influence a language's prosodic structure. The purpose of this paper is therefore to outline an area where we believe this type of analysis is possible and to suggest possible strategies for linguistically and computationally modelling the parameters which affect the assignment of 'focal' versus 'non-focal' tone accents.

Focal and non-focal tone accents in English and Swedish

Although English and Swedish are both Germanic languages, their prosodic systems are considerably different. In American English, the neutral focal tone accent is variously represented as H* (Pierrehumbert 1980) or H*L

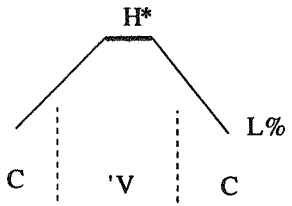


Figure 1a. English neutral/focal (phrase final) tone accent: H* L%

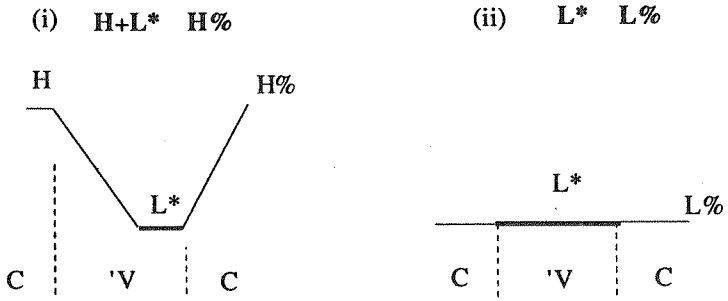


Figure 1b. English non-focal (phrase-final) tone accents.

(Gussenhoven 1984) depending on whether the L is considered to be an integral part of the tone accent or a phrase boundary marker; the H* is associated with the last third of the sonorant part of the stressed syllable (Horne 1987). This tone accent can be realized on both prefocal and focal words within an intonational phrase (Horne 1991). In post-focal position, however, this tonal realization is not found. After focus within an intonational phrase, remaining words are regularly assigned L* tone accents on their stressed syllables instead of H* tone accents; thus, postfocal words are often said to be 'deaccented', where deaccented refers to the lack of a H* tone accent. A L* tone accent, however, is not the only tone-accent realized on non-focal words. A H+L* tone accent, i.e. an 'early peak placement' (see Kohler 1987a, 1987b, Pierrehumbert & Steele 1989, Pierrehumbert & Hirschberg 1990) is also sometimes realized on given information, where the H is linked to the beginning of the stressed syllable and the L* to the latter part of the vowel in the stressed syllable. A following H phrase accent is characteristic of this tone accent pattern which has also been termed 'referring tone' (Brazil 1985). The different tone accent types are schematically illustrated in Figures 1a and 1b (% = 'phrase boundary').

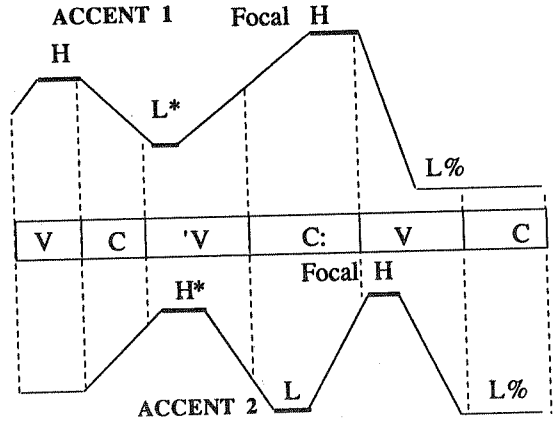


Figure 2a. Schematic representation of Swedish word accents. The association of the starred tone with the stressed syllable is critical. The other associations are only approximate (see Bruce 1977, 1987).

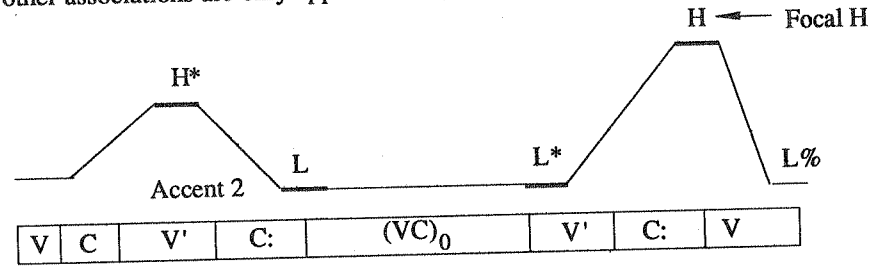


Figure 2b. Accent patterning in Swedish compound words. Association of the starred tone to the stressed syllables is critical. The other associations are approximate and can vary (see Bruce 1977, 1987)

The accenting of given information is implemented somewhat differently in Swedish, since Swedish is a language with two lexical accents which are phonetically stable and realized both in new and given contexts. These are termed Accent 1, or Acute accent which can be represented as HL*, and Accent 2, or Grave accent, which can be represented as H*L (Bruce 1977, 1987). When focused, however, the accents are followed by a H tone, viz. HL*H (Accent 1) and H*LH (Accent 2). In addition to the absence of the final H tone, a non-focal Accent 1 or 2 is usually characterized by a relatively narrower register than a focal one. Even compound words are

characterized by the absence of a final H tone in non-focal contexts, i.e. the compound word accent pattern H*L...L*H, where H* is associated with the first stressed syllable and L* to the last stressed syllable becomes H*L...L* in non-focal contexts. In Figures 2a and 2b (adapted from Bruce 1977) are schematized Swedish focal and non-focal accent patterns in simple and compound words

New/given distinction

Coreferential lexical NP's

Identical morphemes. One important, known factor in conditioning tone accent assignment in texts is the 'new' vs 'given' status of lexical items. In English, for example, given information following the last new lexical item in an intonational phrase is often assigned a non-focal L* instead of a H* tone accent. For example, the word *Saabs* in (1a) would be deaccented because it constitutes 'given' (non-focal) information. The situation is analogous in Swedish, as is illustrated in the corresponding sentence in (1b), where *Saabar* 'Saabs' is also assigned a 'non-focal' accent (i.e. without the focal H) since it was mentioned previously. (In the examples, accented words are written in capital letters and coreferential relations are indicated using the subscript 'i'):

(1) a. MARIE thinks we should INVEST in a SAAB$_i$, but I must ADMIT that I REALLY don't LIKE Saabs$_i$.

b. MARIE tycker att vi borde INVESTERA i en SAAB$_i$, men jag måste ERKÄNNA att jag FAKTISKT inte tycker OM Saabar$_i$.

Keeping track of this type of textual coreference where a lexical item is repeated in a situation of anaphora has been implemented in some TTS systems by having a stack of roots of lexical items mentioned in a portion of the text which is updated at certain fixed intervals, e.g. at paragraph boundaries (Hirschberg 1990).

Coreferential non-identical lexical items: Identity of sense relationships. In order to attain a more complete analysis of textual givenness, however, it is essential to be able to account for several other coreferential strategies that can be used in situations of anaphora and which trigger non-focal tone accents in a similar way as does the repetition of a lexical item. These include the identity of sense relationships known as synonymy, hyponymy,

and part/whole relationships. For example, in (2), the word *dog/hund* (in what follows, we will give English and Swedish examples simultaneously, separated by a slash (/)) constitutes given information because there exists a relationship of hyponymy between it and the word *dachshund/tax*, that is to say, a semantic hierarchy where one term is included in the definition of another. In this case, *dachshund/tax* is included in the superordinate set of *dogs/hundar* and is therefore classified as a hyponym of *dog/hund*. When the superordinate term occurs in a text following the word *dachshund/tax*, it is treated as coreferential to it just as the anaphoric pronoun *it/den* would be and is accordingly assigned a non-focal accent.

(2) a. My SON wants a DACHSHUND$_i$, but I'm not SURE he's OLD enough to take CARE of a dog$_i$.

b. Min SON vill ha en TAX$_i$, men jag är inte SÄKER på att han är GAMMAL nog att ta HAND om en hund$_i$.

In addition to hyponymy, other identity of sense relationships can be observed to express textual givenness (see Allerton 1978, Lyons 1977). Synonymy is one of these, as illustrated in (3):

(3) a. INGVAR thinks that a TAX-RISE is POSSIBLE$_i$, but CARL says that such a measure is not at ALL feasible$_i$.

b. INGVAR tror att en SKATTEHÖJNING är MÖJLIG$_i$, men CARL säger att en sådan åtgärd inte är på något SÄTT genomförbar$_i$.

In (3a), *possible* and *feasible* are synonyms, thus explaining the non-focal status of the latter word at the end of the sentence. The same goes for the Swedish correspondences in (3b), *möjlig* and *genomförbar*. Since *genomförbar* is a synonym of *möjlig*, it is thus contextually coreferent to it and is assigned a non-focal accent when it follows *möjlig*.

A third lexical relationship that is sometimes involved in expressing anaphoric relationships and thus capable of triggering non-focal accent patterns is 'part/whole' relationships. These resemble hyponymy relations in that a word referring to a part of an object (more specific term) can be replaced by a word denoting the whole object (more general term) in a later part of a text. This latter word then receives a non-focal accent as illustrated in (4):

(4) a. My HARD DISK$_i$ CRASHED again this morning. It's got to be the WORST computer$_i$ I've ever HAD!

b. Min HÅRDDISK$_i$ KRASCHADE igen imorse. Det måste vara den SÄMSTA dator$_i$ jag någonsin HAFT!

Note that in the case of hyponymy and part/whole relations, the identity relation is non-symmetrical, so that it is essential that the more general term follows the more specific term in order for the latter to be interpreted as an anaphor to the former and to be assigned a non-focal accent (see Allerton 1978). Placing the more specific term (e.g. *tax*) after the more general term (e.g. *dog*) does not trigger a non-focal accent, as the inappropriate accentual patterns in (5) illustrate:

(5) a. *My SON wants a DOG, but I'm not sure he's OLD enough to take CARE of a dachshund.

b. *Min SON vill ha en HUND, men jag är inte SÄKER på att han är GAMMAL nog att ta HAND om en tax.

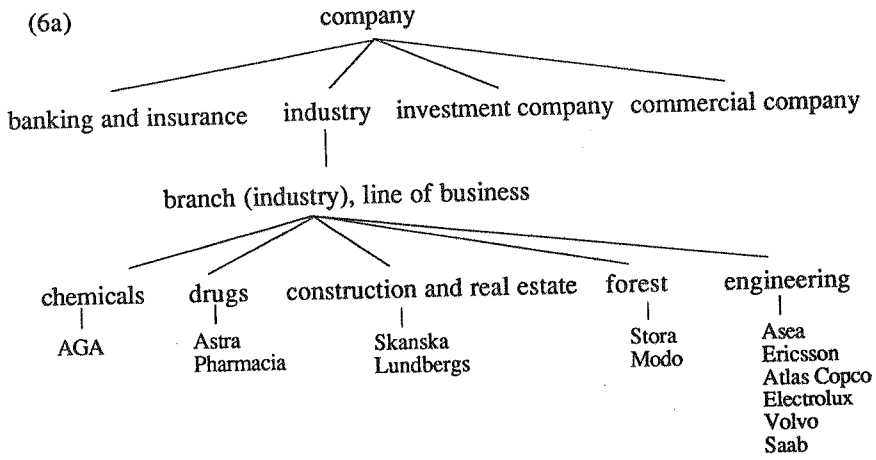
One reason why TTS systems have not implemented detailed semantic rule systems for keeping track of the above mentioned identity of sense relationships is that in order to be able to handle unrestricted texts where one has no specific knowledge of the context, the amount of structuring that would be required and the corresponding computational time involved in searching through all the possible semantic relationships would be horrendous. To take just one rather simple illustrative example, the word *disk/skiva* is polysemous, i.e. it has several meanings and consequently many semantic associations to many other different words depending on the context in which they are used. In the context of physiology, for example, it refers to a part of the human body; in the context of computers, it refers to a component used for storing information; in the context of phonographs, it also refers to an object for storing information, in this case, sound; however, in this context, it is related to other words such as *record/platta*, *single/singel*, *album/LP-skiva* that a computer *disk* is not associated with at all. Thus, because of this semantic ambiguity, in order to decide whether the word *disk/skiva* is coreferent with some other lexical item that was previously mentioned in a text, it would be necessary to search through all the possible semantic relationships that all the meanings of *disk/skiva* can have with other lexical items. Modelling this kind of encyclopedic

knowledge is not currently feasible. Consequently, in order to develop a tractable computational model for structuring lexical information, it seems more reasonable to attempt to represent knowledge in a more limited domain where the ambiguities associated with polysemy can in most cases be avoided.

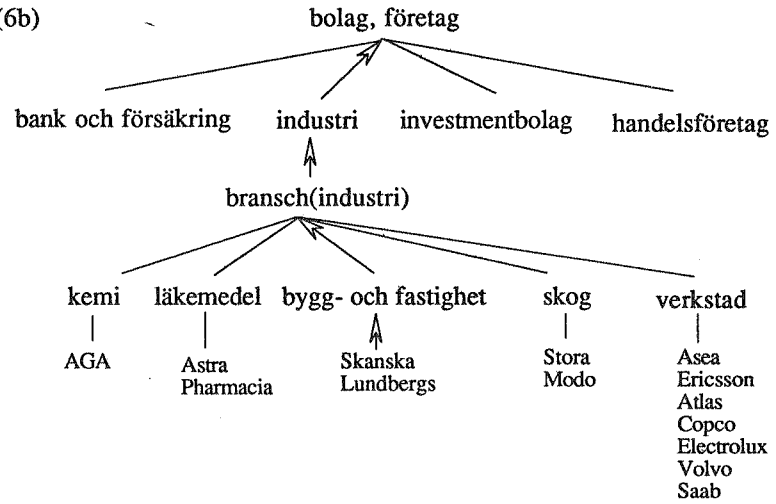
Stock market reports

One domain that lends itself to study in this respect is texts dealing with developments in the stock market. The stock market is a very well-defined world involving buying and selling transactions in stocks and bonds. The reports that appear daily in the newspapers for example deal with changes that the system undergoes during a given day. One important semantic hierarchy that is relevant when determining textual givenness in stock market reports is that structuring the relations between individual companies and their superordinate structures, such as the partial hierarchies shown in (6):

(6a)



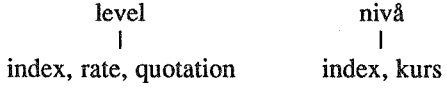
(6b)



This hierarchy represents facts such as the following: that Astra is a drug/läkemedel company, and that drugs/läkemedel, construction and real estate/bygg- och fastighet, and forest/skog are all branches/branscher of industrial companies/industri. Thus, one would expect that for example, after having mentioned Astra shares, one could refer to them anaphorically by calling them shares in drugs/läkemedel.

Another semantic relation that is present in the stock market world is that realized in expressions for *stocks/aktier* and *bonds/obligationer*. These can be referred to anaphorically by the superordinate term *paper/papper*. One can also see a rather rich inventory of terms used to cite the value of a given stock on the market; there would seem to be a relation of synonymy between the terms *index/index*, *rate/kurs*, and *quotation*/n.a., and a relationship of hyponymy between them and the more general term *level/nivå*. These can be represented as in (7):

(7)



We will represent synonymy between different terms on a given hierarchical level by grouping them together under a single branch.

One can also expect that certain concepts in the stock market world are 'situationally given' (i.e. taken for granted) for all speaker/hearers (see Chafe 1974, Firbas 1979). A potential candidate for this status is the word

denoting the particular currency used for a given country's stock exchange reports. In Sweden, the *crown/krona* can be expected to have this status. If it does, then one could mark it in the stock market lexicon as contextually given ([-new]); consequently, it would never be assigned a focal accent in speech. The same goes for situational information such as the day and year of the particular stock market report since this is given in the newspaper.

Data

In order to gain more detailed knowledge of the lexical structure of stock market texts, we have made a preliminary lexical and prosodic investigation of a newspaper text taken from *Sydsvenska Dagbladet*. The text, together with an English translation, was read and recorded by male native speakers of Standard Swedish and American English, respectively. This particular text has also been the subject of study in the SWETRA (automatic translation) project at the Dept. of Linguistics (Sigurd 1990). A portion of the text is reproduced below. Coreference relationships have been marked by means of identical bold subscripts. The symbol '+' is used in the Swedish text to show the boundary between component morphemes in a compound.

Stockholm's$_{a}$ stock exchange$_{b}$ general index$_{c}$ closed$_{d}$ on Thursday$_{e}$ at 858.8, a marginal increase$_{f}$ of 0.02 percent$_{g}$ compared with Wednesday's$_{h}$ closing$_{d}$ index$_{c}$. The rate$_{c}$ development$_{i}$ during the day$_{e}$ was described as irregular.

Rate$_{c}$ increases$_{f}$ in AGA$_{j}$ and Astra$_{k}$ made the trade index$_{c}$ for chemicals$_{j}$ and drugs$_{k}$ to be the best line of business$_{j,k}$ with an increase$_{f}$ of 1.6 percent$_{g}$.

The losers were the remaining industrial and trading companies$_{j,k}$, which fell by 2.6 and 1.9 percent$_{g}$, respectively.

The buying rates$_{c}$ rose in 52 companies$_{j,k}$, fell in 80, while 189 remained at Wednesday's$_{h}$ closing$_{d}$ levels$_{c}$.

The rate$_{c}$ development$_{i}$ for OTC$_{l}$-companies$_{j,k}$ and O$_{m}$-listed companies$_{j,k}$ was very weak. The OTC$_{l}$ index$_{c}$ fell 1.2 percent$_{g}$, and the O$_{m}$-index$_{c}$ went back 1.3 percent$_{g}$.

After a receding opening trading$_{n}$ on Thursday$_{e}$ at Stockholm's$_{a}$ Exchange$_{b}$, the very positive semiannual reports$_{o}$ from AGA$_{j}$ and Ericsson got the rate$_{c}$ development$_{i}$ to turn around.

Stockholms$_{a}$ fond+börs$_{b}$ general+index$_{c}$ slutade$_{d}$ på torsdagen$_{e}$ på 858,8, en uppgång$_{f}$ med marginella 0,02 procent$_{g}$ jämfört med onsdagens$_{h}$ slut$_{d}$+index$_{c}$. Kurs$_{c}$+utvecklingen$_{i}$ över dagen$_{e}$ betecknades som oregelbunden.

Kurs$_{c}$+stegringar$_{f}$ i AGA$_{j}$ och Astra$_{k}$ fick bransch$_{j,k}$+index$_{c}$ för kemi$_{j}$- och läkemedel$_{k}$ att bli bästa bransch$_{j,k}$ med en uppgång$_{f}$ på 1,6 procent$_{g}$.

Förlorare var övrig industri$_{j,k}$ och handels+företag, som föll med 2,6 respektive 1,9 procent$_{g}$.

Köp+kurserna$_{c}$ steg i 52 bolag$_{j,k}$, föll i 80 medan 189 låg kvar på onsdagens$_{h}$ slut$_{d}$+nivåer$_{c}$.

Kurs$_{c}$+utvecklingen$_{i}$ för OTC$_{l}$-bolagen$_{j,k}$ och O$_{m}$-listade företag$_{j,k}$ var mycket svag. OTC$_{l}$-index$_{c}$ föll med 1,2 och O$_{m}$-index$_{c}$ gick tillbaka med 1,3 procent$_{g}$.

Efter en vikande inledande handel$_{n}$ på torsdagens$_{e}$ Stockholms$_{a}$+börs$_{b}$, så fick de mycket positiva delårs+rapporterna$_{o}$ från AGA$_{j}$ och Ericsson kurs$_{c}$+utvecklingen$_{i}$ att vända.

The decline during the morning then recovered successively due to rising quotations$_c$ in market-leading paper$_p$. Sales during the session$_n$ were few and rose to just under 234 million crowns$_q$, of which almost 50% represented sales in Astra$_k$, Ericsson, and Bilspedition.

After a temporary depression during Wednesday$_h$ in connection with profit winning, it was time again on Thursday$_e$ for Astra$_k$ shares$_p$ to rise during a brisk trading session$_n$. The best quotation$_c$ was noted for the free B-share$_p$, which ended at 520 crowns$_q$, 15 crowns$_q$ above Wednesdays$_h$ closing$_d$ quotation$_c$. Astra$_k$ presented its nine-month$_r$ report$_o$ on Wednesday$_h$, which shows a profit increase of 35 percent compared with the same period$_r$ last year. Owing to the continued successes for the gastric ulcer preparation Losec, Astra$_k$ has been able to write up its prognosis for the whole of 1990 to nearly 2.5 billion crowns$_q$.

Tillbakagången under förmiddagen återhämtades sedan successivt tack vare stigande kurser$_c$ i marknads+ledande papper$_p$. Omsättningen under sessionen$_n$ var liten och gick endast upp till knappt 234 miljoner kronor$_q$, varav nära 50 procent utgjorde handel i Astra$_k$, Ericsson och Bilspedition.

Efter en tillfällig svacka under onsdagen$_h$, i samband med vinst+hemtagningar, var det på torsdagen$_e$ åter dags för Astra$_k$+aktierna$_p$ att stiga under en livlig handel$_n$. Bästa kurs$_c$+utvecklingen$_i$ noterades för den fria B-aktien$_p$, som slutade till 520 kronor$_q$, 15 kronor$_q$ över onsdagens$_h$ sista$_d$ kurs$_c$. Astra$_k$ presenterade under onsdagen$_h$ sin nio+månaders$_r$+rapport$_o$, som visar en vinst+ökning på 34 procent$_g$, jämfört med samma period$_r$ i fjol. Tack vare de fortsatta framgångarna för magsårs+medlet Losec har Astra$_k$ också kunnat skriva upp sin prognos för hela 1990 till närmare 2,5 miljarder kronor$_q$.

Observed tone-accent patterns in data

As one can see from examining the text, there are a considerable number of anaphoric relations present. The second and third paragraphs contain instances of the superordinate structure in (6).

In this text, one observes that the specific companies *AGA* and *Astra* are referred to later in the text by *chemicals/kemi* and *drugs/läkemedel*. These latter terms are then later referred to by the expression *line of business/bransch*, and in the final sentence, the expression *industrial company/industri* is used to refer back to the individual branch industries. Thus, following the argumentation above, we would expect that *AGA* and *Astra* would be assigned focal accents, since they constitute new information; however, we would not expect that the superordinate terms that are used to refer back to them anaphorically, i.e. *chemicals/kemi*, *drugs/läkemedel*, *line of business/branch*, *industrial company/industri* would be assigned focal accents. This is, in fact, what one finds when one examines Fo patterns on the words under discussion. In Figure 3 are presented the American English intonational patterns associated with *AGA and Astra*, *chemicals and drugs*, and *line of business*, and in Figure 4 the corresponding Swedish intonation patterns.

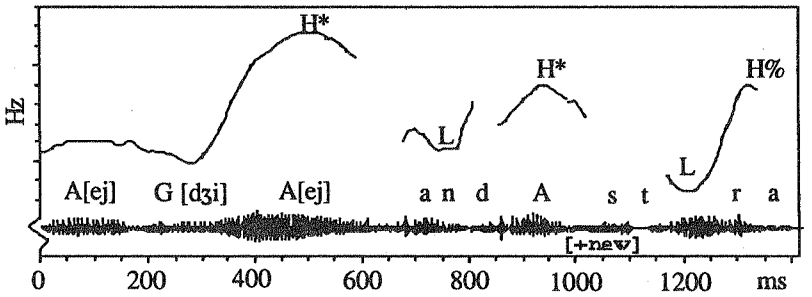


Figure 3a. Fo pattern on the phrase *AGA and Astra*. Both *AGA* and *Astra* constitute [+new] information and are assigned H*L tone accents. The phrase boundary is marked by a H boundary (%) tone.

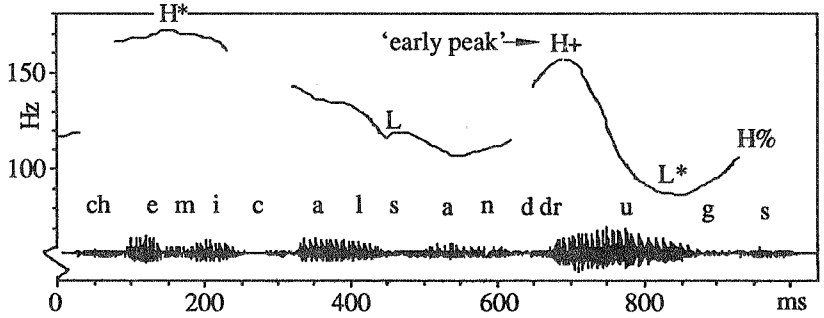


Figure 3b. Fo pattern on the phrase *chemicals and drugs*. The 'nuclear' accent on *drugs* ([-new]) is realized as a H+L* ('early peak') tone accent followed by a H % boundary tone.

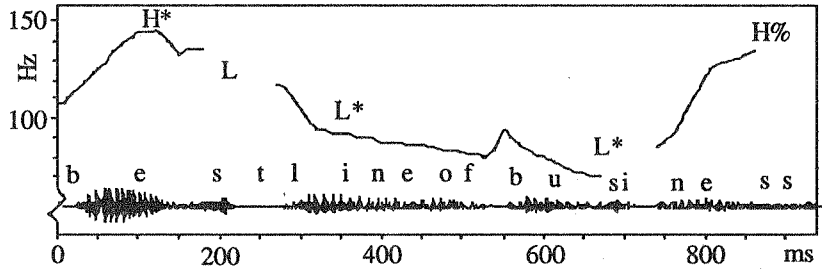


Figure 3c. Fo pattern on *best line of business* where *best* constitutes [+new] information and *line of business* constitutes 'given' ([-new]) information. Best is accordingly assigned a H* L tone accent, whereas *line of business* is 'deaccented', i.e. assigned a L* tone accent.

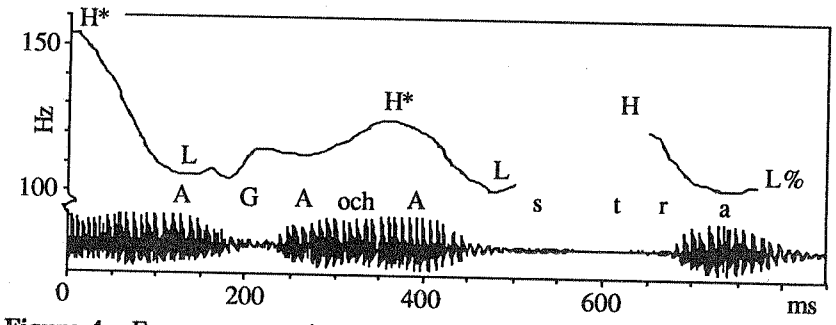


Figure 4a. Fo pattern associated with the phrase *AGA och Astra* ([+new]). Both words are assigned Accent 2 (H*L). *Astra* is also assigned a 'focal' H in its final syllable.

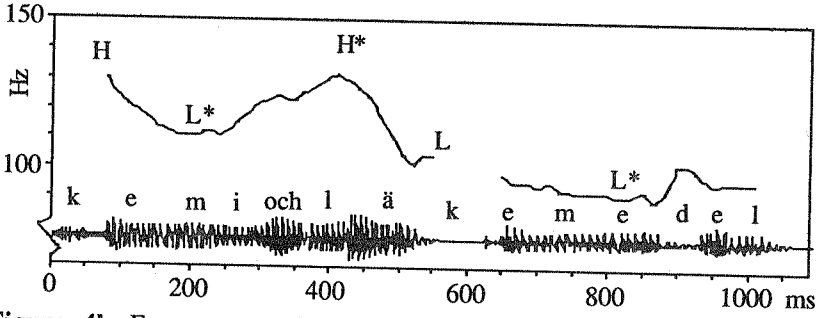


Figure 4b. Fo pattern on *kemi och läkemedel* 'chemicals and drugs' ([-new] information). *Kemi* exhibits a non-focal HL* accent pattern as does the compound word *läkemedel*, i.e. it has an Accent 2 (H*L) pattern on the first stressed syllable, and a L* on the last stressed syllable.

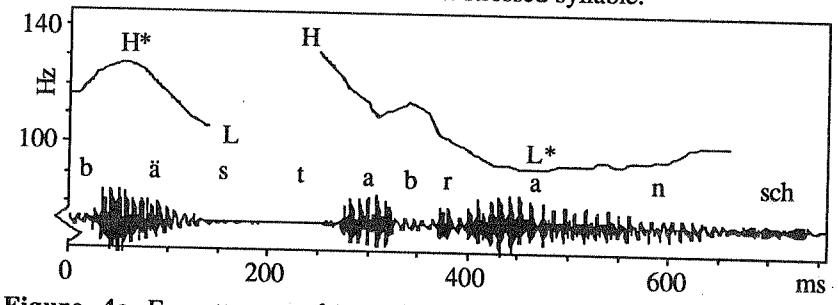


Figure 4c. Fo pattern on *bästa bransch* 'best branch' where *bästa* constitutes [+new] information and *bransch* [-new] information. *Bästa*, an Acc. 2 (H*L) word, is thus associated with a focal H in its final syllable. *Bransch*, an Acc.1 word is not, however, associated with a focal H.

Computational modelling of new vs. given information

As we have established in the previous discussion the type of tone accent pattern assigned a lexical item depends to a large extent on whether it functions as new or given information.

Finding new information

To find out what is given information in a text is a function of what has been mentioned before. On the word level there are two cases which can be separated. In the first case the current *word* or a derivate from it has been mentioned before. This case corresponds to the finding and matching of character strings.

In the second case the *meaning* of the current word has been mentioned before. This second case is much more difficult because it does not only involve a matching of the surface forms but also a matching of some abstract meaning connected to that form. The meaning of the word can be stored in a lexicon in some form which is easy to handle and easy to retrieve and/or update. The two cases will be elaborated upon in the following two sections.

Finding strings which have been mentioned before. When looking for new information in a text, one must be able to check whether a pair of words have the same stem, i.e. if their surface forms have the same referent. It may seem to be a simple problem to match surface forms but it can be very difficult indeed if we are looking for a 100% accurate algorithm. On the other hand it is quite easy to do with a method that applies in most cases. Some examples of the function we would like to have are presented in the rules in (8):

(8) same_stem(company, companies) → **true**
same_stem(company, compare) → **false**
same_stem(closed, closing) → **true**
same_stem(closed, closet) → **false**

This matching can be accomplished for example in English by comparing the words from left to right until we find a mismatch. When a mismatch is found then we say that this position is the beginning of the word endings of the words being compared. When we have found the word endings (which can possibly be a null string), their existence is checked for in a dictionary of possible endings.

The question 'do word A and B have the same stem?' can be reformulated as the weaker question 'if AE is the end of word A and BE is the end of word B and the remaining part of both words is the same, is it then true that both AE and BE are word endings?'. In our material there seems to be a good probability that the weaker question is sufficient to answer the original question with good accuracy:

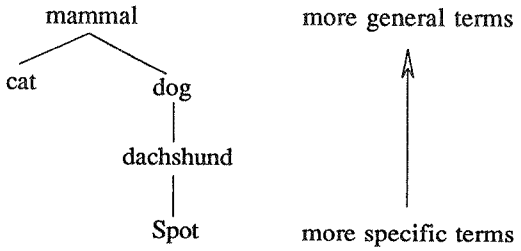
'same_stem(company, companies)' leads to a check on *y* and *ies* which is true in English. 'same_stem(company, compare)' leads to a check on *ny* and *re* which are not good endings in English. Our table of endings would include neither *ny* nor *re*. Now we can say that *company* and *companies* probably contain the same lexical meaning but *company* and *compare* do not.

If we have mentioned *company* before then it is not *new* information to mention either *company* or *companies* again. This approach will also work well for Swedish because Swedish like English uses suffixes to mark grammatical derivations. The tables of endings are of course different in the two languages.

Finding meanings which have been mentioned before. Finding out if the meaning of a word has been mentioned previously in a text is a more difficult problem than finding out if a string of letters has occurred previously. It is, however, possible to keep track of meanings in limited text, just because it is possible to predict which concepts and superordinate expressions will probably be mentioned.

Superordinate hierarchies such as those presented above in (6) can be described as directed trees where the daughters of each node define a more specific level. New and given information are related to these trees by the fact that it is in some sense new information to make a specification, but to make a generalisation is a revelation of given information (in the context of our described hierarchies):

(9)



In the context of (9), *dog* is a specification of *mammal* and therefore subordinate to *mammal*. In the other direction *dog* is not a specification of *dachshund*. To describe these relations, a predicate is required which has the following effect:

1) specification_of(dog, mammal) → **true**
2) specification_of(dog, Spot) → **false**

To summarize the above discussions: If any word has the same stem as a previously found word or it is a semantic specification of a previously found word then it is classified as given. In all other cases we can presume that it *is* new information.

Restricting the scope of new information

Some of us humans constantly forget what has previously been mentioned. It can be argued that there is a *need* to refresh old information once in a while and treat old information as if it were new information.

The difficult problem is to know when 'old information' becomes 'new information' again. A helpful way out could be to allow the machine to forget' that something has been mentioned before. But how should this forgetting take place?

1) After a certain number of words (the list length is limited)
2) Randomly with a certain probability
3) Other strategies, e.g. after the end of paragraphs (Hirschberg 1990).

In longer texts, it will definitely be necessary to forget that words have been mentioned for computational reasons. The time required to insert or find an element in a list is proportional to the length of the list and therefore, without forgetting, the machine would run slower and slower as the length of the list increases. Another reason for keeping the old information only for a limited time is that after a while we would have found almost all words and therefore almost nothing will be new information any more.

Possible lexical conditioning of given information. As mentioned above, some information is contextually/situationally given in a text. That is to say, some words can be so common in the context or have such a general meaning that they are never assigned tone-accent patterns associated with new information. In our sample material we have the example of the word

for Swedish currency which can be said to be given (taken for granted) information in the context of the Swedish stock market; in our text, the word *crowns/kronor* is moreover the most frequently used lexical item. Thus we can make the prediction that words denoting a given country's currency are never assigned a focal tone accent in stock market texts. This suggests that in some cases the distinction between new and given information can be marked directly in the lexicon. An interesting finding with respect to the data presented here is that the distinction between new and given information seems to be connected to the specific/general dimension in this restricted domain. The more general a word is the more likely it is that it functions as given information. The more specific a word is the more likely it is that it provides new information. This distinction is reflected not only at the lexical level but also at the phrase level. Thus if a noun phrase consists of an Adjective+Noun, the Adjective functions as a semantic specifier of the Noun and is more likely to be assigned the focal accent than the Noun. For example, note the focally accented words in the following passage of the data:

> The decline during the morning then recovered SUCCESSIVELY due to RISING quotations$_{\mathbf{c}}$ in MARKET-LEADING paper$_{\mathbf{p}}$.
>
> Tillbakagången under förmiddagen återhämtades sedan SUCCESSIVT tack vare STIGANDE kurser$_{\mathbf{c}}$ i MARKNADS-LEDANDE papper$_{\mathbf{p}}$.

Due to the fact that *quotations/kurser* and *paper/papper* can be assumed to be contextually coreferent with preceding information, since *quotations/kurser* is synonymous with *index*, and *paper/papper* is superordinate to *shares/aktier*, the new information is expressed by the Adjective specifiers *rising/stigande* and *market-leading/marknadsledande*. Thus one could propose a rule for focal prominence by saying that if the head noun in a Noun Phrase is found to constitute given information, then the focal accent falls on the attribute (see Horne 1987). This rule assumes, however, an active search process through the text which could be time-consuming. If it were the case, however, in restricted texts such as stock market texts, that specifiers such as adjectives were most likely used only when the head word was given, then one could have a lexical rule that marked Adjectives for carrying focal accents. Analysis of further data is necessary, however, in order to confirm this hypothesis.

Conclusion

In this paper, we have proposed a method for analysis of restricted texts that attempts to account for factors that trigger the assignment of focal versus nonfocal tone accents in texts. Specifically, we have examined situations of anaphora where lexical words are used to refer back to their antecedents. The coreference relations that trigger the assignment of non-focal accents in these contexts are explainable as resulting either from morphological identity or from semantic identity of sense relations such as synonymy, hyponymy and part-whole relations. It has been suggested that in restricted texts, much of this information can be modelled computationally. Some information can be specified in the lexicon, for example marking of certain words as situationally given information or as semantically specific words with respect to this particular domain. An interface to a TTS system incorporating the lexical information discussed here would allow the system to generate more natural tone accent patterns than those that are currently available. More specifically, the H+L* and L* tone accents for English could be generated on given information; moreover, in the case of Swedish, assignment of non-focal word and compound word accent contours could be more accurately modelled.

References

Allerton, D. 1978. 'The notion of 'givenness' and its relations to presupposition and to theme'. *Lingua* 44, 133-168.

Brazil, D. 1985. 'Phonology: Intonation in discourse'. In T.A. van Dijk, (ed.) *Handbook of discourse analysis* Vol. 2, 57-75. London: Academic Press.

Bruce, G. 1977. *Swedish word accents in sentence perspective*. Lund: Gleerup.

Bruce, G. 1987. 'How floating is focal accent?'. In K. Gregersen & H. Basbøll, (eds.), *Nordic prosody IV: Papers from a symposium*, 41-49. Odense: Odense Univ. Press.

Chafe, W. L. 1974. 'Language and consciousness'. *Language* 50, 111-133.

Firbas, J.. 1979. 'Post-intonation-centre prosodic shade in the modern English clause'. In S. Greenbaum, G. Leech and J. Svartvik, (eds.), *Studies in English linguistics for Randolph Quirk*, 125-133, London: Longman.

Gussenhoven, C. 1984. *On the grammar and semantics of sentence accents*. Dordrecht: Foris.

Hirschberg, J. 1990. 'Using discourse context to guide pitch accent decisions in synthetic speech'. *Proceedings of the ESCA workshop on speech synthesis (Autrans, France)*, 181-184.

Horne, M. 1987. 'Towards a quantified focus-based model for synthesizing English sentence stress'. *Lingua* 75, 25-54.

Horne, M. 1991. 'Why do speakers accent "given" information?'. *Proceedings of Eurospeech 91:2nd European conference on speech communication and technology*, Vol. 3, 1279-1282.

Kohler, K. J. 1987a. 'The linguistic functions of Fo peaks'. *Proceedings XIth Int. Cong. of Phonetic Sciences, Tallinn, Estonia.* Vol. 3, 149-152.

Kohler, K. J. 1987b. 'Categorical pitch perception'. *Proceedings XIth Int. Cong. of Phonetic Sciences, Tallinn, Estonia.* Vol. 5, 331-333.

Lyons, J. 1977. *Semantics*. Cambridge: Cambridge University Press.

Pierrehumbert, J. B. 1980. *The phonology and phonetics of English intonation*. Unpublished dissertation, MIT. (Distributed by Indiana University Linguistics Club, 1988).

Pierrehumbert, J. B. & J. Hirschberg. 1990. The meaning of intonation contours in the interpretation of discourse. In P.R. Cohen, J. Morgan & M. E. Pollack, *Intentions in communication*, 271-311. Cambridge, Mass.: MIT Press.

Pierrehumbert, J. B. & S.A. Steele. 1989. Categories of tonal alignment in English. *Phonetica* 46, 181-196.

Sigurd, B. 1990. *SWETRA - a multilanguage translation system. Praktisk lingvistik* 14 (Dept. of Ling., Lund University).