Appendix B: Annotation and data preparation details

B.1 Technical annotation details

The files were first transcribed in ELAN (Sloetjes & Wittenburg, 2008). The transcriptions were then exported to Praat (Boersma & Weenink, 2018), which was used to adjust the word segmentations. The adjusted segmentations were then re-imported in ELAN prior to doing the annotations.

For the annotation of BAs, Praat was used for inspecting the f_0 contour, but annotations were made in the ELAN file (i.e., Praat and ELAN were used simultaneously while annotating BAs).

B.2 Differences between data subset 1 and 2 with respect to annotation principles

Subset 1 (the older part of the corpus, 30 stories) was fully annotated for BA, HB, and EB by three annotators (Annotator 1-3), while subset 2 (additional 30 files) was annotated by two new annotators (Annotator 4 and 5), one at a time. This discrepancy is explained by our annotation history. We started off with three annotators because we had the resources, we needed to develop and discuss annotation principles, and we had no knowledge on how reliable these annotations would be. When reliability turned out acceptable (Section 2.2.3), we decided to proceed with one annotator per file. At the same time, there was a need for more efficiency, because we drastically increased our dataset (from 31 to 142 files; note, however, that only 60 of these files are included in the present study, as we added labels of f_0 landmarks for these 60 files only).

In addition, principles for HB and EB annotations differed between the two subsets, as follows: For subset 1, annotations of HB and EB were done with full access to the audio- and video channels, as well as a display of the word segmentations (see Section 2.2.2 for a motivation of this procedure). For the additional news stories, HB were annotated as before, while EB were annotated in a second step, by another annotator, without access to the audio channel and without prior listening: In one half the subset, Annotator 4 labelled HB, while Annotator 5 labelled EB (without access to the audio) and then BA; the other half of the subset was then labelled with switched roles. This new routine was introduced as we strived for gestural annotations being made with access to the video only. However, for the HB annotations, this was deemed difficult due to the large number or HBs in the corpus and the principal difficulty of afterwards assigning HB annotations to words (see Sections B.4 and B.5). For the EB annotations, this was judged much more feasible as our previous annotations had revealed rather few instances of EB in this kind of data. As a result, the annotations are slightly inconsistent, but our comparison of HB annotations made with and without access to the audio suggests no serious consequences of this inconsistency (see Sections 2.2.2 and 2.2.3).

B.3 Correction of BA annotations

Cases with an original disagreement between the three annotators for subset 1 (e.g., where only two annotators had labelled a BA) were revisited and discussed by the authors, resulting in occasional corrections. Mistakes were only corrected when obvious: Either there was a clearly identifiable big-accent rise present, but the word was not labelled as BA; or, a word was labelled as BA, while no rise was identifiable. Such mistakes are explainable, as annotators occasionally were misled by their auditory impression of the prominence of a word, although the instruction was to look for BA-rises in the f_0. However, a word may sound very prominent, although it lacks a BA; likewise, a word may be produced with a BA, but still not stand out from its environment (especially in news reading, which contains numerous BAs per sentence). Likewise, annotations from subset 2 (which had been labelled by only one annotator per file) were corrected where necessary.

B.4 Converting time-aligned HB annotations to word-based HB annotations

The time-aligned HB annotations typically did not overlap with only one particular word, but typically with two (or more, when short functions words were involved). Thus, we needed a criterion for assigning the time-aligned annotations to concurrent words. Defining this criterion might seem theoretically challenging, as it would seem to require knowledge about how prominence-lending head movements are synchronized with the word that would most likely benefit most from the movement, in terms of increased prominence. Practically, however, this task was rather trivial in the context of the present study. The purpose of the comparison with the original annotations was only to test whether the same movements would be recognized when annotators have access to the audio, compared to when only video is available. We were not interested in the timing precision of the annotations. Therefore, we assigned the time-aligned HB annotations to words as follows. We compared the time-aligned HB annotations with the original categorical annotations (to be more precise: with the final classification described in Section 2.2.4): In case one of the words that overlapped with the time-aligned HB annotations was associated with an original, categorical HB-annotation, we assigned the time-aligned HB annotation to that word. In such a case, the original and the new time-aligned annotation were thus counted as equivalent in the comparison, as it is obvious that the same movement had been seen by the new annotator and the original annotator(s). However, in case there was no HB-label in the original annotation, we simply assigned the time-aligned HB-annotation to the word that showed the largest temporal overlap with the HB-internal. (Note that it would not make a difference for the present purpose if the criterion had been defined in another way; it was only important to compile a record of all HBs that were either seen in both annotation settings or in only one of the settings.)

There was a complication, though. The annotator of the time-aligned annotations was allowed to distinguish between ‘simple’ (monodirectional) and ‘complex’ annotations. The comparison with the categorical annotations revealed that, oftentimes, a ‘complex’ label corresponded to two original HB-annotations. In order to define an objective criterion for converting these ‘complex’-annotations into categorical, word-based annotations, we generally assigned a HB to two of the words that overlapped with the complex time-aligned HB, even if there was only one HB in the original annotation.

Technically, the conversion of the time-aligned HB annotations into word-based annotation was performed as follows: We added the original word annotations to the ELAN-files containing the new time-aligned HB annotations, and manually assigned each time-aligned HB to the word that it was to be associated with according to the criteria described above. We used a separate tier for that purpose. The content of this tier (from all 8 files), augmented with the original HB-annotations, was the input for the calculation of a κ-value (Section 2.2.3).

B.5 More on the classification of movements and accents as multimodal prominence constellations (MMP)

Words were annotated for belonging to either of the three conditions of interest, either BA, BA + HB, or BA + HB + EB. The information on cooccurrences of BA, HB, and EB was indirectly available in the ELAN-annotations, however, spread out over several tiers, from different annotators. This classification was made after export of the annotations to Praat. In order simplify data processing after extraction of f_0 values, new labels were defined and labeled on new Praat tiers (“1” = BA, “2” = BA + HB, “3” = BA + HB + EB).

As mentioned in Section 2.2.4, this classification was not entirely trivial. The problem is that head and/or eyebrow beat annotations do not necessarily occur in the word they most likely refer to, but sometimes earlier (see Ambrazaitis & House, 2017). In particular, different annotators (in subset 1) may have assigned the same movement to different words (see Section 2.2.2). For instance, two annotators might have recognized a HB on word x, while the third annotator has recognized a HB on word x-1, which is, obviously, the same movement that was seen by the other two annotators. However, in a number of cases, two or all three annotators recognized the movement on the preceding word. It could be a HB, an EB, or both HB and EB that was recognized on the one or two words preceding the accented word (see Figure 6 for an example). This happened in 45 cases in total (8% of the data), where a majority of cases (32) involved an EB that was recognized on a preceding word (either only the EB in 22, or both the HB and the EB in 10 cases); the remaining 13 cases were of type BA + HB (the HB being annotated on a preceding word).

Even in such cases, we would argue that there are good reasons to assume that the HB and the BA (and the EB, if present) form a cluster, because the general pattern observed in our previous studies is that HBs occur in connection with accented words, and EBs in connection with HBs. Furthermore, for EBs a tendency of slightly preceding the HB has been observed (House et al., 2017; see also Flecha-Garcia, 2010, who showed that eyebrow movements precede pitch accents).

We therefore classified a word as ‘BA + HB’ or ‘BA + HB + EB’ not only if the movement(s) occur(s) on the word, but also up to two words in advance. The threshold of two words was determined *ad hoc* during the classification process: In a majority of cases, a one-word threshold would have seemed reasonable, but there were cases where a stretch of very short function words preceded the BA-accented word, and head and/or eyebrow annotations were spread out during the 1-2 preceding words. The threshold was hence set to two words in order to capture such cases.

Two examples (see also Figure B1 for a third example):
(1) Two succeeding words have been annotated as BA and HB; each label is provided by at least two of three annotators. Both words are classified as ‘BA + HB.’
(2) A word has received a BA-label by all three annotators, and a HB-label by one annotator. Another annotator has put a HB label on the preceding word (which is not labelled BA). Hence, considering even the preceding word, two annotators (a majority) have annotated a HB. In addition, all three annotators have seen an EB, but one of them on the preceding word (x-1), and the other two on word x-2. The word is classified as ‘BA + HB + EB.’

B.6 Details concerning the labelling of f_0 landmarks

The actual f_0 labels were more complex than those shown in Figure 2 in the main paper, as they also included a number denoting the word accent category (see Figure B1): The initial H was labelled either H1 or H2, the following L either L1 or L2, and the final H was labelled FH1 or FH2 (where the “F” marked the H as representing the big-accent [or ‘Focal’] H).

For A2 words, the determination of the three landmarks was generally easy, as both maxima are typically clearly realized, H2 early in the stressed syllable, and FH2 either in the post-stress or in the secondary stress (in compounds). The minimum was then identified as the lowest f_0 in between the two maxima. However, care was taken to exclude f_0 values that obviously represent micro-prosodic effects. To this end, two principles were generally applied: First, f_0 labels were never placed within (voiced) obstruents (and generally not in unvoiced segments), even though the actual f_0 minimum was often found within such segments; instead, the f_0 minimum was placed in the segment before or after the obstruent (wherever f_0 was lowest). In addition, a distance of about 2 oscillations from the segment boundary was maintained. Second, f_0 maxima were often observed on or right after a segment boundary, especially after a voiceless obstruent (as in a /t/ + vowel syllable onset), were f_0 often falls from a high level as an effect of consonant articulation. To avoid measuring these micro-prosodic peaks, again, a distance of about 2 oscillations from the segment boundary was maintained in such cases. This approximate distance measure was defined *ad hoc*, after recognizing that it would usually succeed in excluding the most extreme f_0 values caused by consonant articulation.

For A1 words, the big-accent H-landmark (FH1) was generally easily determined, too, as it was defined as the maximum reached during, or sometime slightly after, the stressed syllable. However, determining the accentual H and L was often more problematic, and its feasibility was rather depending on the context. The accentual fall in A1 is expected to happen, roughly, from the pre-stress to the stressed syllable, with a L-tone early in the stressed syllable. However, the big-accent H can be realized early, right from syllable onset, too. In order to capture a rise at all, in such cases, the minimum should be measured in the pre-stress. Another complication was that in some cases, a local f_0 maximum was determinable in the pre-stress (often only a small peak), but often not. Hence, to define a general rule for annotating the landmarks H1 and L1, L1 was defined as the f_0 minimum preceding the FH1 landmark either within the stressed syllable or in the pre-stress (but not earlier). The H1 landmark was defined as the f_0 maximum preceding L1 within the pre-stress; if there was no maximum (e.g., if the L1 had been set at the initial segment boundary of the pre-stress), no H1 was labelled. These rules provide reliability, while the validity of the accentual fall in A1 measured this way can be questioned. In particular, it is likely that (small) falls are measured (when H1 was present) that might not be of true relevance for the accent (see the discussion in Section 4.1.2 in the main paper). However, the alternative to these rules would be to search for H1-maxima even earlier, probably reducing reliability of labelling and further reducing the validity of the measure. Finally, the same principles for avoiding the measurement of segmental effects were adhered to as defined above for A2 words.

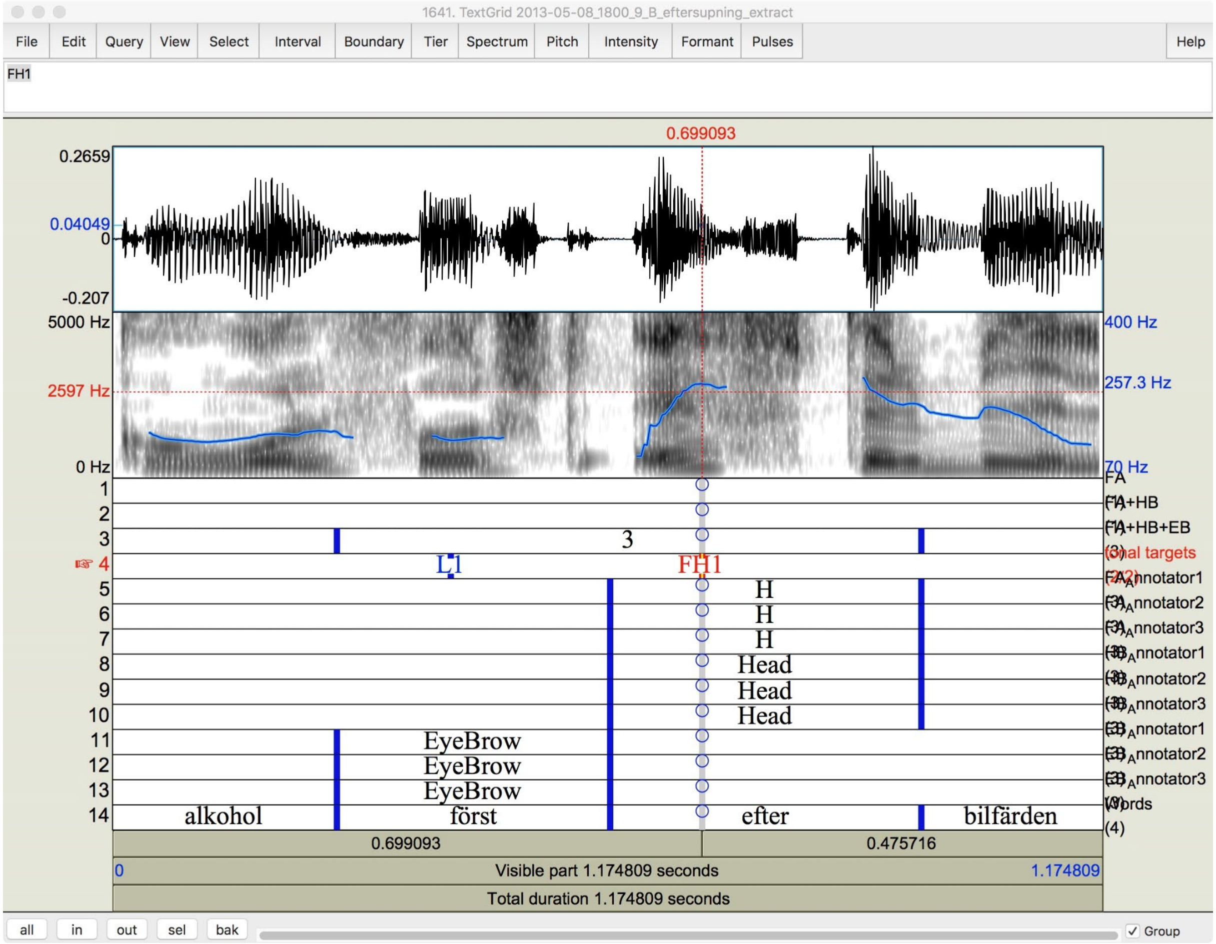

Figure B1. Screenshot of a Praat window illustrating the process of annotating f_o landmarks (B.6) and the classification of patterns (B.5). The example shows the phrase *alkohol först efter bilfärden* 'alcohol first after the car ride' with a BA on the initially-stressed Accent 1 word *efter* 'after.' In this case, the L tone was identified in the pre-stress syllable (= the monosyllabic word *först*), and no initial H tone (which would have been labelled 'H1') could be identified within the pre-stress syllable (as the preceding maximum falls into the pre-pre-tress syllable -*hol*); see text for explanations. Note that a BA is referred to as 'FA' ('focal accent') in the figure and labelled using the label 'H'. The tonal landmark capturing the f_o maximum is labelled 'FH1' for 'H tone of the focal accent in Accent 1.' This figure also illustrates how gesture annotations sometimes precede the accented word. In this case all three annotators agreed that the HB occurred on the accented word, while the EB occurred on the preceding word (see Section B.5). The word was classified as a case of BA + HB + EB, coded using the label '3.'

The f_o landmarks were labelled using ProsodyPro (Xu, 2013). The script was also used to manually correct the temporal placement of glottal pulses, where necessary, which were automatically determined by Praat. These pulses were then used to calculate f_o (see Section B.7). f_o annotations and pulse corrections were performed by a research assistant and then checked and corrected by the first author.

B.7 Data extraction and additional tagging

f_0 landmarks, together with the word transcriptions and the summarizing labels of accent/gesture clusters defined above ("1," "2," "3," for BA, BA + HB, and BA + HB + EB) were extracted automatically using a custom-written Praat script. In order to avoid unnecessary f_0 analysis errors, the f_0 calculation provided by ProsodyPro was used, which is performed in the time-domain based on the manually corrected 'pulses' (see Section 2.2.5). The script also applies a smoothing algorithm removing minor spikes from f_0 curves.

The output was reorganized as a datafile usable in R and augmented with information on speakers (name and sex), original file name, and a classification of the word as either A1 or A2 (this information was embedded in the f_0 labels). The table also contains a classification of the words as either 'simplex' (one lexical stress) or 'compound' (two lexical stresses). In addition, a column was added where the content of the news story was coded, as the material contains occasional repetitions of the same news story read at several times during a day. Also, a news story can be a continuation of another, adding new information to the same case. All stories that clearly referred to the same case (repetitions or continuations) were encoded using the same label. This coding resulted in 34 different stories among the 60 files. Finally, another column was added grouping the stories roughly into topics. This classification is not central for the present analysis, but is mentioned here because the Topic factor was included as a random-effects factor in our mixed models (see Section 3.2 in the main paper). The resulting datafile is included in the supplementary materials (Appendix C).

References

Ambrazaitis, G., & House, D. (2017a). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication, 95*, 100-113. doi:10.1016/j.specom.2017.08.008

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer*. Computer program. http://www.praat.org/.

Flecha-García, M.L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication, 52*(6), 542–554. doi:10.1016/j.specom.2009.12.003

House, D., Ambrazaitis, G., Alexanderson, S., Ewald, O., & Kelterer, A. (2017). Temporal organization of eyebrow beats, head beats and syllables in multimodal signaling of prominence. In *International Conference on Multimodal Communication: Developing New Theories and Methods*. Osnabrück, Germany.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7–10). Aix-en-Provence, France.