SIGNSPEAK

Scientific understanding and vision-based technological development for continuous sign language recognition and translation

Grant Agreement Number 231424

Small or medium-scale focused research project (STREP)

Project start date:  1 April 2009
Project duration:  36 months

Deliverable D. 2.1. A scientific paper on the marking of sentence boundaries in NGT

Authors: Ellen Ormel & Onno Crasborn
Radboud University Nijmegen, Centre for Language Studies

Dissemination Level: Public
Ormel & Crasborn (subm.), The prosodic correlates of sentences in signed languages and their relation to language technology: a literature review. Sign Language Studies.

Content

1. Introduction ......................................................................................................................3
2. Literature review .............................................................................................................5
   2.1 Prosody and sentences in spoken languages .................................................................5
   2.2 Prosody in sign languages ............................................................................................6
   2.3 Prosody at sentence boundaries in Sign Languages ....................................................8
       2.3.1 Studies of eye blinks .................................................................................................9
       2.3.2 Studies of multiple sentence boundary cues ..........................................................10
       2.3.3 Conclusion: combinations of phonetic cues to prosodic boundaries ......................16
   2.4 Automatic detection of sentence boundaries in spoken languages ............................16
   2.5 Automatic detection of sentence boundaries in signed languages ..............................16
3. Suggestions for two types of empirical studies Suggestions for two types of empirical studies .....................................................................................................................18
   3.1 Study 1: New tests of human segmentation of signed sentences ...............................18
   3.2 Study 2: Experimental tests of human vs. machine segmentation .............................19
4. Discussion ........................................................................................................................21
References ................................................................................................................................22
1. Introduction

The identification of signed sentences in a larger stream of discourse is not a trivial task. Linguists have to decide in their analysis what counts as predicate, arguments and other syntactic material, and distinguish main clauses from embedded material. The semantic and syntactic analysis of signed languages is an ongoing matter of research and debate (e.g. Liddell, 1980, 2003; Johnston, 1991; Engberg-Pedersen, 1993; Neidle et al., 2000; Sandler & Lillo-Martin, 2006). How lexical and syntactic material actually appear to the viewer, i.e. the phonetic form of the language is mediated by the phonological level of organisation. Phonological forms above the syllable have commonly been dubbed ‘prosody’, just as the rhythmic and melodic properties of spoken languages (Sandler, 1999ab, Nespor & Sandler, 1999). While many authors argue that the overall design of the grammar of signed languages shows large similarities to spoken language organisation, the phonetic substance of signed and spoken languages are very different (Brentari & Crossley, 2002; Sandler, to appear). The phonetic correlates of rhythm and intonation in signed languages consist of non-manual activities and modifications of manual phonological material. As at the syntactic level of organisation, it is not self-evident how sentence-like units in signed languages can be identified on the basis of phonetic properties of the prosodic form. The same would seem to be true of sign language users: perceivers cannot simply watch for one or two specific facial cues in order to segment the signing stream in sentences or large prosodic domains that are the phonological correlate of such syntactic units. At the same time, it is well-known from psycholinguistic studies on spoken language that prosody does help in the perception and recognition of spoken language (Gerken, 1996, Cutler et al., 1997, Frazier et al., 2006); it is likely that this is also the case for signed language processing.

In a recent theme issue of the journal Sign Language & Linguistics on the unit ‘sentence’ in signed languages, both syntactic and prosodic perspectives were taken on the sentence in empirical studies of three signed languages (Crasborn, 2007). These all (implicitly or explicitly) subscribed to the conception of prosody as being related to, but not a direct expression of, syntactic structure (cf. Sandler, to appear; Crasborn, 2010). Thus, phonetic events (whether in the face or on the hands) indirectly point to phonological structure, just as F0 (Fundamental frequency) and duration in spoken language are reflections of tones and prosodic groupings in speech (Pierrehumbert, 1980, Selkirk, 1984, Ladd 1996, Gussenhoven, 2004). Phonological domains such as the utterance and the intonational phrase are related to syntactic structure, but the actual phrasing will depend in part on performance factors like speaking rate (Selkirk, 1984, Nespor & Vogel, 1986). Thus, the same syntactic string of words can be articulated in different ways, with different phrasing, showing more Intonational Phrases when it is articulated very slowly than when it is realised at high speed. The overall model of the relation between syntax and phonetics is thus highly similar to what Shattuck-Hufnagel & Turk (1996) sketch for speech. One of the possible models they conceive is presented in Figure 1. An alternative they suggest is a joint phonological component for prosody and segmental phenomena). While there is little consensus on the presence of a unit ‘segment’ in the phonology of signs, there surely is a sub-syllabic level of organisation that makes up the core of the phonological representation of lexical items (consisting of one vs. two-handedness, selected fingers and their configuration, orientation, a place of articulation, and movement properties) (Sandler, 1989; Brentari, 1998; Crasborn, 2001; van der Kooij, 2002; van der Kooij & Crasborn, 2008).
In the present paper, we do not contribute new evidence to either perspective on the sentence, but we aim to present a literature review of prosodic evidence for large prosodic domains that can be equated with a syntactic unit ‘clause’ or ‘sentence’ (section 2). This paper was produced in the context of a European project on automatic sign language recognition and translation (SignSpeak; http://www.signspeak.eu), in which (after recognition of lexical items) the stream of signs in a video recording has to be segmented in some way by software in order to be able to translate sign language to the written form of a spoken language. This translation process is tied to sentences in the output language, in the case of the SignSpeak project these are in written Dutch. For that purpose, it is a relevant question to ask whether sentence units are in some way visible in the video recordings of signed discourse. It is not crucial per se whether the units in question are actually sentences or sentence parts (clauses). For that reason, in this paper we will not address the topic of syntactic structure (and by consequence, neither the question how specific phonological domains are generated from the syntax). In addition, we will use the term ‘sentence’ and ‘clause’ fairly loosely, aiming to refer to a high syntactic domain that can receive a translation in a full Dutch sentence. In reality, this will of course depend on the actual syntactic form in Sign Language of the Netherlands, in our case.

On the basis of the various ideas and views in the literature, we describe two types of possible studies that can contribute to our understanding of prosodic cues to sentence boundaries (section 3). First of all, refinements are suggested to the perception studies by Nicodemus (2009), Hansen and Heßmann (2007), Hochgesang (2009), and Fenlon et al. (2007), in which the intuitions of signers and non-signers are elicited on break points in stretches of discourse. In addition, we sketch how the use of new techniques from visual signal processing can be exploited to detect salient events in a video recording of signing. This can facilitate the analysis of larger data sets than would have been possible by phonetic transcription. Finally, in section 4 we end with a brief discussion on how knowledge of the prosody of signed
languages can be employed for language technology, such as the automatic translation from signed to spoken language.

2. Literature review

In order to interpret a sentence adequately and to decide what pieces of information should be part of the same sentence and what belongs to a different sentence, several sources are used, such as syntactic, semantic, and discourse information. In addition, prosodic information provides additional support to interpret sentences in spoken as well as in signed languages. Many studies of spoken languages showed that the form of those languages is intrinsically determined by prosody (Cutler, Dahan, & van Donselaar, 1997). In a literature review on prosody in the comprehension of spoken language, Cutler et al. (1997) described that the term prosody is used in different ways: ‘From one extreme those who maintain an abstract definition not necessarily coupled to any statement about realization (“the structure that organizes sounds”) to those who use the term to refer to the realization itself, that is, effectively use it as a synonym for suprasegmental features (“pitch, tempo, loudness, pause”). The majority describe prosody between these two extremes, in using the term to refer to abstract structure coupled to a particular type of realization (“the linguistic structure which determines the supra segmental properties of utterances’)’ (Cutler et al., 1997: 142.). The aim of most studies of prosody is to understand the process of recognition.

In contrast to prosody in spoken languages, prosody in signed languages has only been studied to a limited extent, in particular when it comes to studies of sentence boundaries. First we will describe some studies on prosody in relation to sentences in spoken languages.

2.1 Prosody and sentences in spoken languages

In spoken languages, prosody is a widely discussed topic. Within the research on prosody, boundary signals or features have been the most studied phenomena (Cutler et al., 1997). In a tutorial on prosody, Shattuck-Hufnagel and Turk (1996) explained the characteristics in spoken utterances that written equivalents do not have in terms of patterns of prosodic structure; intonation, timing and variations in segmental implementation. The organization of a spoken utterance is not isomorphic to its morphosyntactic structure. Therefore prosody is a significant issue for (auditory) sentence processing. The same is likely to be true for sign language processing, if one assumes the same general design of the grammar as in Figure 1.

Definitions of prosody have been variable. Definitions can refer to the acoustic parameters, presumably signaling constituent boundaries and prominence: F0, duration, amplitude and segment quality or reduction. Other definitions can refer to the phonological organization of parts into higher-level constituents and to the hierarchies of relative prominence within these constituents. Constituents include, for example, intonational phrases, prosodic phrases, and prosodic words. A third group of definitions combine the phonological aspect of prosody at the higher level organization with the phonetic effects of the organization (e.g., F0, duration, segment quality/reduction). Shattuck-Hufnagel and Turk (1996) use the following working hypothesis: ‘Prosody is both (1) acoustic patterns of F0, duration, amplitude, spectral tilt, and segmental reduction, and their articulatory correlates, that can be best accounted for by reference to higher-level structures, and (2) the higher-level structures that best account for these patterns (p.196). Evidence for prosodic constituents has been based on phonological observations as well as acoustic phonetic measurements.

In addition, evidence has been provided for the psychological reality of prosodic constituents, coming from studies of language behavior. For example, unit monitoring studies have provided evidence for the initial perceptual organization of spoken utterances. Other
evidence on the psychological reality of prosody is based on the comparison between the perception of pauses or interruption points within constituents versus pauses at constituent boundaries. To conclude, Shattuck-Hufnagel and Turk (1996) provided four helpful hints to support further research on auditory sentence processing studies that may form a starting point for looking at signed utterances: ‘1. Since prosody can’t be predicted from text, specify the prosody of stimulus utterances as it was really produced; 2. Since prosody can’t be read off the signal alone, inform acoustic measurements by perceptual transcription of the prosodic structure of target utterances; 3. Consider interpretation of results in terms of prosodic as well as morphosyntactic structure; 4. Define those terms’ (p.241). We will take up especially the second point in our suggestion for future approaches in sections 3 and 4.

One example of work concerning the psychological reality of prosodic constituents was carried out by Bögels, Schriefers, Vonk, Chwill, and Kerkhofs (2009) who suggested that prosodic information could be sufficient to determine the syntactic analysis of a sentence. In other words, prosody not only provides some additional support for the interpretation of sentences, but it directs the syntactic analyses. Cutler et al. (1997) explained that the presence of prosodic information cueing a boundary can influence syntactic analyses of listeners; however, they suggested that the evidence for this influence is not robust as yet. Prosodic information does not always cue syntactic information directly. In other words, it has been known that prosody is informative towards syntactic structures, but whether it has a conclusive role is presently open for discussion. Moreover, many studies have shown that the available prosodic cues in spoken languages are not always exploited by the listener (Cutler et al., 1997). Furthermore, work on prosody in relation to sentence boundaries by Carlson, Clifton, & Frazier (2001) suggested that the interpretation of a prosodic sentence boundary is related to the existence and relative size of other prosodic boundaries in the sentence. This suggestion is in contrast to suggestions that boundaries have a confined effect independent of context. These different findings already show that conclusions about the specific role of prosody in spoken language processing are still refutable.

Research on prosody in sign language has started even more recently, and many questions remain unanswered, at this point. The study of sentence boundaries in relation to prosody in sign language is one of the areas that have been studied to a limited extent only. The role of prosody in sign language will be discussed next, followed by a review of a number of studies on prosody at sentence boundaries in sign languages. Subsequently, several recent developments are described on automatic detection of sentence boundaries in spoken and signed language.

2.2 Prosody in sign languages

Similar to spoken languages, prosody can also be observed in sign languages, providing subtle and meaningful supplements to sentences (e.g., Dach kokovsky & Sandler, 2009; Wilbur, 2000). Several studies on prosodic constituents have been conducted in sign languages. Already in 1991, Allen, Wilbur & Schick studied the rhythmic structuring of sign language, in their case ASL, by three groups of signers: ASL fluent deaf signers, ASL fluent adult hearing children, and non-signing hearing adults. The participants were asked to tap a metal stick in time to the rhythm of signed narratives. The participants tapped in each of the three groups in a rhythmical way. In the paper, Allen et al. (1991) explain than non-native signers are often detected by the lack of rhythm is their signing. Some differences were found between the rhythmic identification patterns by the hearing non-signers and by the deaf signers, which showed that the rhythm of ASL is only observed fully when participants have a good grasp of sign language. One difference was that non-signers were not able to make the distinction between primary and secondary stress. Another difference was that adult hearing children of deaf adults (CODAs) showed much greater variability within the group than the other two
groups did, suggesting that the rhythm of spoken English may have interfered for this group of CODAs.

The temporal aspects of signs which might function as prosodic elements were analyzed for Swiss-German Sign Language by Boyes Braem (1999), in an attempt to determine what makes the signing of native signers seemingly more rhythmic than the signing of late learners. For that reason, three early signers were compared to three late signers. Two kinds of rhythmic patterns (typically referring to beat or stress) were found in the study, which may have resulted in the observation of a sign as being rhythmic. One rhythmic pattern was the temporal balancing of syntactic phrases. Temporal balancing refers to phenomenon that the final phrase is produced with approximately the same duration as that of the preceding phrase. No differences were found between early and late learners of sign language. The second kind of pattern was a regular side-to-side movement of the torso which appears to mark larger parts of certain types of discourse phonetically. Early learners showed this pattern more than late learners.

Prosodic cues can occur on the manual articulators (hands and arms) and nonmanual articulators (face, head, and body). Nonmanuals typically also add semantic information to the manual signs. Nonmanual markers include head position, body position, eye brow and forehead, eye gaze, nose position, and mouth, tongue and cheek actions. As Wilbur (2000, p. 237) described: ‘Nonmanual markers are integral components of the ASL intonation system, performing many of the same functions in the signed modality that pitch performs in the spoken modality’. In a number of studies, nonmanuals were studied to gain insight in prosodic boundaries. In general, nonmanuals cue either the end of phrases (boundary markers) or their extent (domain markers). The moment of onset and offset of the markers provide important information about their function (Wilbur, 2000). Wilbur also found that the large degree of layering in American Sign Language is a linguistic adjustment to the modality. Layering serves prosodic and pragmatic purposes. The most apparent layering is that of the simultaneous production of nonmanual markings with manual signs.

One of the most frequently mentioned boundary markers are eye blinks. Already in 1978, Baker & Padden first introduced the importance of eye blinks in sign language research, in particular the inhibited periodic blinks at the end of intonational phrases (see also Wilbur, 1994). There are also voluntary blinks that co-occur with lexical signs and perform a semantic and/or prosodic function of emphasis, assertion, or stress. The finding that voluntary blinks can be followed immediately by periodic blinks at the end of the constituents (slow and fast blinks) resulted in a study on the different functions of blinks. Also head nods serve multiple functions, e.g., focus, marking emphasis, assertions, and/or existence. In other situations, head nods seem to function as edge marking, similar to blinks. Wilbur concluded that head thrusts do not seem to have a prosodic or syntactic function, but rather a semantic function, similar to some of the head nods; indicating the genuineness of or the speaker’s commitment to the statement in a clause. Head thrusts are articulated with the lower jaw thrust forwards, and these occur on the last sign in some specific clauses (e.g. ‘if’ and ‘when’ clauses in ASL).

Several nonmanual articulators can have a function as domain markers: eye brow position, eye gaze, head tilt, negative headshake, body leans, and body shift. These domain markers often appear to begin at the start of a constituent and end when the constituent ends. As they are articulated independently from the manual string of signs and each articulator can maintain a specific position, these positions can be kept constant for a certain time span. It is this phonetic affordance of the visual modality that makes the presence of domain markers specifically prominent and their roles quite diverse. Although tone in spoken language can also be raised or lowered for a certain domain, the many factors influencing F0 in speech make it quite hard to ‘mark’ domains for a perceiver in a consistent way. For some nonmanual articulators in signed languages, it is slowly becoming apparent that there too, there are more
influences on their phonetic appearance than was initially thought (see for example the study of De Vos, van der Kooij & Crasborn (2009) on eye brows in Sign Language of the Netherlands).

Three eye brow positions are often distinguished: raised, lowered, and neutral (but see e.g. Baker-Shenk (1983) and De Vos et al. (2009) on the appearance of more detailed patterns). Raised brows occur on a number of apparently unrelated structures. Neutral brows occur on assertions, and furrowed brows occur with WH-questions and embedded WH-complements in some signed languages (Zeshan 2006). Prosodic modification of brow behavior can be observed as an effect of increased signing rate. The number of brow raises decreases as a result of fast signing rate without changing the syntactic structure (the number of constituents accurately marked by brow raise stay similar). Eye gaze as well as head tilts express agreement information. These can co-occur with each other and with different brow positions. Negative headshakes have a grammatical instead of an affective function. The interaction of negative headshakes with different head positions such as head tilt and head nod require further analysis, although it is to be expected that these are unable to coexist, given the semantic distinct reasons to use negative headshakes versus tilts and nods. Body leans can be side-to-side leans for prosodic purposes (as described by Boyes-Braem in 1994), indicating other signers or different locative or temporal situations, but body leans can also move forward and backward (see also Wilbur and Patschke, 1998). The different leans can indicate prosodic emphasis on particular lexical items; semantic categories of inclusion ‘even’ and exclusion ‘only’; and type of contrastive focus (Wilbur, 2000, p. 235). Van der Kooij, Crasborn & Emmerik (2006) found similar functions for Sign Language of the Netherlands. They found support for a multi-layered functioning of meanings: forward leaning was associated with a different set of meanings, i.e. involvement, eagerness (lexical level), highlighting of information (subject focus), affirmation of sentence content, and belief; whereas backward leaning was associated with dislike, disgust, rejection (lexical level), highlighting of information (object focus), negation of sentence content, and disbelief. The authors emphasize that it was not always easy to distinguish linguistic and paralinguistic functions.

Body leans as well as body shift may be differentiated by their prosodic behavior, whereby body leans are restricted to signs and body shifts are restricted to left-edge markers within a certain (undetermined) domain (Wilbur, 2000). Finally, the nose has various contributions to sign language. Nose wrinkling can function as an affective marker but also as a contextual discourse marker. According to Wilbur, there are probably other nose functions that are unknown as yet.

Part of the work on prosody in spoken as well as in signed languages has focused on segmentation and sentence boundaries. In the following part, summarise a series of studies that have specific impact on the identification of clause and sentence boundaries in signed languages.

2.3 Prosody at sentence boundaries in Sign Languages

In spoken languages, length, pitch, spectral properties and voice quality (and possibly others) together constitute the prosodic form of language. Cues from these domains can be used by the listener to segment a string of sounds into words or larger units. According to Cutler et al. (1997) there has only been little discussion of the relative time-course of prosodic processing in sentence comprehension (in spoken languages, but even more true for signed languages). In the following overview and subsequent thoughts for future directions in section 3, we attempt to achieve more insight in prosodic processing in sign languages, in particular in relation to segmentation of signed discourse into clause or sentence units. We start with a review of the
literature on eye blinks, which is frequently discussed, and then move on to studies which aim to look at various cues in combination.

2.3.1 Studies of eye blinks
In a study by Wilbur (1994), signers of American Sign Language were studied for the inhibited periodic eye blinking. Wilbur showed that the eye blinks of ASL signers were sensitive to syntactic structure. These findings provide insight into how intonational information shows in a signed language, information which is carried by pitch in spoken languages. Wilbur made a broad distinction between nonmanual markers carried on the lower part of the face and nonmanual markers on the upper face and head/body. The first were analysed as performing a lexical semantic function, whereas the latter ones carry grammatical and prosodic function. The nonmanuals that mark phrasal boundaries are blinks, head nods, change of head/body, and pauses. Wilbur (1994) distinguished three basic types of eye blinks: reflexive blinks (not studied), involuntary or periodic blinks and voluntary blinks.

The results in the 1994 study showed that signers typically blink at the end of Intonational Phrases (right edge of ungoverned maximal projection). Moreover, voluntary blinks are longer in duration and have greater amplitude than involuntary blinks. Voluntary blinks occur at lexical signs and involuntary/periodic blinks function as boundary markers. Four different functions can be identified for blinks at boundaries: the marking of syntactic phrases, prosodic phrases (intonational phrases), discourse units, or narrative units.

Sze (2004) performed a study to examine the relationship between blinks, syntactic boundaries and intonational phrasing in Hong Kong Sign Language (HKSL). Wilbur’s findings on voluntary blinks (marking emphasis, assertion, or stress) and involuntary blinks (marking intonational boundaries) in ASL were evaluated empirically in HKSL. This classification is insufficient to explain all data in the study on HKSL. Blinks in HKSL were bearing a high correlation with head movement and gaze changes. Moreover, some blinks were boundary sensitive but they could co-occur with syntactic boundaries of constituents equivalent to or smaller than a clause.

Wilbur made a distinction between lexical blinks and boundary blinks. Lexical blinks are voluntary and occur simultaneously with the lexical item, whereas boundary blinks are involuntary and occur at intonational phrases. According to Sze’s (2004) analysis, some findings by Wilbur were less clear in her data. These differences related to the exact measurement of the duration of a blink, the duration of a sign, and determining whether a blink is produced voluntary or involuntary. Whereas in Wilbur’s study, 90% of the boundary blinks fall right on the intonational phrase boundaries, this was only true for 57% in HKSL. Moreover, many of the blinks (30%) in HKSL data co-occur with the last sign in a sentence, which seem to function as boundary marking, but according to Wilbur’s description, these would have to be interpreted as a lexical blink. Third, Sze described that some blinks are accompanied with another blink at the same sign. To account for these differences, Sze proposes a new classification system of types of blinks: 1. Physiologically induced; 2. Boundary sensitive; 3. Related to head movement/gaze change, unrelated to syntactic boundaries; 4. Voluntary/lexically related; 5. Associated with hesitations or false starts. Type 1 and 5 do not have a linguistic function in the sense of being related to a specific grammatical structure, although they are likely to have a function in perception in being related to the performance phenomena related to false starts or hesitations. Sze concludes by suggesting that changes in head movement may, in fact, be better than blinks to serve as clues for intonational phrase boundaries, given that blinks often co-occur with head movements. Moreover, it is uncertain whether the addressee is aware of blinks, whereas head movements are more directly observable.
In a study of prosodic domains in ASL, Brentari & Crossley (2002) have used an extreme interpretation of Wilbur’s findings, by adopting eye blinks in their methodology as the basic dividing marker between Intonational Phrases. On the basis of that division, further studies of manual and nonmanual markers of Phonological Phrases were sought. It appears that they did not distinguish between different types of eye blinks, nor accounted for the 10% of IP boundaries that were not accompanied by an eye blink in Wilbur’s study on ASL.

Crasborn, van der Kooij, & Emmerik (2004) performed a small similar study concerning the prosodic role of eye blinks. Frequencies and locations were measured of blinks in different types of data (monologues versus question-answer pairs), distinguishing between short and long blinks. The results indicated that a distinction seems to be present between shorter and longer blinks. Longer blinks seemed to serve functions similar to other ‘eye aperture’ functions: wide eyes and squinted eyes. Wide eyes express surprise, disgust and emphasis, whereas squinted eyes express fear and shared information. Closed eyes may express disgust and counter-assertion, and may be related to reference. With Sze (2004), the authors emphasize that the frequency of involuntary blinks and possibly also the duration of other types of blinks are likely to be influenced by the physical context, including the air humidity, amount of dust in the environment, and the temperature. This makes it hard to compare results across studies, and also findings from different recording sessions.

One open question based on the literature is what counts as a long blink and a short blink. The results also showed that blinks occurred at many locations, and it is an open question whether they all relate to linguistic elements. There does not appear to be a strict mapping of one articulator to one function in the eye opening parameter. Another open question is whether sign language perceivers actually can and do perceive blinks, including the brief involuntary blinks in normal interaction. It is clear there is no clear view for any sign language on the role of eye blinks in the perception of prosodic structure.

Finally, we would like to note that for studies of eye blinks (as in the other literature on nonmanuals), the difference between grammar and prosody is awkward if one adopts an overall model of language structure as in Figure 1. Nonmanual forms are often characterized in phrases like ‘Head shake is the grammatical marker of negation’. However, grammatical features (such as negation) have a prosodic phonological shape (such as head shake), and it is this phonological shape that can interact with the prosodic structure and with other phonological features. What we can actually see in a video recording is not a phonological shape but one specific phonetic instance of more general phonological categories. In other words, what can be observed in the phonetic appearance (a video recording, but also a kinematic recording of movement trajectories for example) is a cue that is somehow related to the phonological form that may have a semantic or grammatical function. It is likely that most instances of the phrase ‘nonmanual (grammatical) marker’ refer to the combination of a semantic value and its phonological shape.

In addition to phonological forms that are linked to a semantic or syntactic feature, there may appear to be prosodic phonological features that do not have any semantic or grammatical content, but only have the function to signal prosodic boundaries. If one would compare inhibited periodic eye blinks as similar to boundary tones in spoken languages, for example, then even these still relate to the grammar in marking the prosodic domains Intonational Phrase for example which is derived from the syntactic structure of the sentence.

2.3.2 Studies of multiple sentence boundary cues
The finding that there is non-isomorphism between syntactic and prosodic constituents indicates that the requirement for rhythmic structure forms an independent property or phonological organization (see also Sandler, 2006). Nespor & Sandler (1999) defined the basic research theme of finding out whether prosodic patterns are exclusively present in
spoken languages, or whether they also occur in sign languages. The latter would make it more likely that prosody is a universal property of human language irrespective of modality. Nespor & Sandler (1999) presented initial evidence showing that ISL sentences indeed do have separate prosodic constituents such as phonological phrases and, at a higher level in the prosodic hierarchy, Intonational Phrases (see also Sandler, 2006).

In their study, Nespor and Sandler found that four markers almost always occur at phonological phrase boundaries in ISL: 1. Reduplication (reiteration of the sign); 2. Hold (freezing the signing hand or hands in their shape and position at the end of the sign); 3. Pause (relaxing the hands briefly); and 4. Separate facial articulation. Hold, reduplication and pause may all belong to the same phenomenon of lengthening given that these markers are never used simultaneously (see e.g. Byrd, Krivokapic & Lee, 2006, on the distribution of lengthening in spoken language). Moreover, the occurrence of an increased number of repetitions is dependent on the lexical sign having a feature [repeated movement]. Each of these specific prosodic cues may be perceptually prominent, such that co-occurrence is unnecessary. However, the modality favors a layering of information, for example for facial articulations such as eyebrows, eyelids, mouth which may be simultaneously layered on one another. The issue of simultaneous occurrence of prosodic cues requires further examination in future studies.

According to Sandler (2006), even more obvious prosodic cues occur at the boundaries of Intonational Phrases compared to Phonological Phrases. Intonational Phrases are at a higher prosodic level than Phonological Phrases. In spoken languages, Intonational Phrases are produced in one breath and begin with a new breath of air. There is a correspondence between eye blinks in sign language and breathing in spoken language in that both are imposed by the physiology of our body and will regularly occur irrespective of whether we are uttering linguistic units or not. It would therefore not come as a surprise if eye blinks would indeed indicate intonational boundaries in sign languages. In addition to blinks, two other characteristics noticeably indicate IP boundaries: changes in head position and major changes in facial expression. During an IP, head position seems to remain constant up to the point of boundary, where it clearly changes, providing a rhythmic cue for those phrases.

Nespor and Sandler suggested themselves that although rhythmic structures were clearly shown, it would be interesting to use larger and more varied corpora, statistical analyses, and experimental studies to confirm to suggested patterns. In addition to providing evidence that prosodic cues form a highly valuable contribution for the detection of sentence boundaries in running sign language videos, one interesting aspect of their findings so far is the apparent ability to define boundaries on the basis of prosodic form, irrespective of content. This in turn is especially promising for the development of automatic sign language translation machines. Aside from the need for more quantitative research in this domain, we would like to add.

In an extensive empirical study, Fenlon, Denmark, Campbell, & Woll (2007) examined whether deaf native signers agreed on the locations of boundaries in the narratives and what cues are used when parsing the narratives. In addition, hearing non-signers were asked to mark sentence boundaries in order to compare the visual cues used in deaf native signers and hearing non-signers. Six native signers of BSL and six non-signers were asked to mark sentence boundaries in two narratives: in BSL and in SSL. Narratives were segmented in real-time, using the ELAN annotation tool (http://www.lat-mpi.eu/tools/elan/). Before assessing participants’ responses, all Intonational Phrase (IP) boundaries in both BSL and SSL narrative were identified using a cue-based approach. Firstly, the occurrence of a blink between signs was used to indicate possible IP boundaries and these boundaries were further verified by the presence of other cues such as pauses and head nods. Following identification, a strict 1.5 second window was applied to all IP boundaries in both signed narratives in which
responses associated with that boundary could occur. Results indicated that the majority of responses from both groups in both narratives fall at IP boundary points (instead of Phonological Phrase boundary or a syntactic boundary).

The results showed that the number of cues present at each boundary varies from 2 to 8 (occurring simultaneously), showing prosodic layering in the sense of Wilbur (2000). Several cues were frequently present such as head rotation, head nods, blinks, and eye brows. Blinks were one of the cues that occurred highly frequently, however, not many blinks occurred at boundaries detected by many signers. Some other cues, such as pauses, drop hands and holds seemed to occur mainly at strong IP boundaries. Because the cues could occur simultaneously and sequentially, it was difficult to detect the cues that were actually used by the participants to segment the narratives in the real-time online segmentation. No relationship was found between the number of cues present at an IP boundary and the identified boundaries. Further, participants seemed to be looking for the same cues in a signed narrative regardless of whether they know the language or not. However, segmentation was more consistent for the deaf signers than for the hearing signers.

The main conclusions that were drawn were that phonetic cues of prosody form reliable indicators of sentence boundaries, while cues from grammar were argued not to be essential for segmentation tasks, as shown by the fast process in real-time. Some IP boundaries are perceptually stronger than others and can even be identified by those who do not know the language. For the cues at those boundaries (pauses, drop hands, holds), language experience seems to play a minor role only. It is therefore likely that the IP boundaries here coincide with even higher boundaries, such as of the prosodic domain Utterance, or an even larger discourse break. As far as the theory of prosodic phonology is concerned, all IP boundaries are equal, even though the phonetic cues of a specific IP boundary might be more prominent than those of another one, depending on context factors for example. Moreover, it was suggested that the many occurrences of head rotations and head movements may be unique to narratives.

We find this type of study to be very valuable. Further studies are needed to verify the findings and confirm them for other languages, and to analyze further any differences in boundary detection between those who are native users of a language and those who are not. In particular, it is questionable whether at normal speed viewing, signs do not make use of their semantic and syntactic processing in determining boundaries, and really rely only on prosodic cues. We come back to this issue in a suggestion for further studies in section 3.

More research is needed to discover exactly how boundaries of identical and different layers differ (e.g., PP vs. IP vs. Utterance), whether in duration, intensity or type of cue marking. Furthermore it would also be interesting to examine the differences and similarities in boundary identification between deaf and hearing participants if monologues or dialogues other than narratives would be used.

Hansen and Heßmann (2007) stated that none of the formal markers, such as blinks, change of gaze, lengthening and transitions, are conclusive in detecting sentence boundaries, which cannot be determined independently from meaning in German Sign Language (DGS). Nevertheless, three cues were argued to be useful indicators in determining sentence boundaries in DGS. A short sample of DSG text was segmented based on a functional analysis called TPAC (Topic, Predication, Adjunct, and Conjunct). The TPAC analysis supports the identification of boundaries of nuclear sentences. The results of this functional analysis were largely in line with, and to a degree refined, results based on intuitive judgements. As a next step, the occurrence of specific manual signs, interactively prominent gestures, head nods, eye blinks, gaze direction, pauses, and transitions were compared to the segments based on the analyses of propositional content to test if prosodic cues of sentence
boundaries occur consistently or exclusively at relevant boundaries, as shown by the TPAC analysis.

The results showed that none of the prosodic cues consistently function as a boundary marker. They stated: “As we will argue, signers recognize sentences by identifying propositional content in the course of a sense-making process that is informed but not determined by such form elements” (p. 146). Nevertheless, temporal adjuncts such as PAST and NOW seemed to point to an antecedent boundary of some sort. Also the palm-up gesture requires attention in future studies on sentence boundaries (see also Crasborn, van der Kooij & Ros 2006). Hansen and Heßmann showed that 60% of the palm-up gestures appeared at a sentence boundary. The remaining 40% referred to multiple further functions of the palm-up gesture, which altogether would appear to make it an inconsistent cue for sentence boundaries by itself. Similarly, head nods with a concluding force may indicate sentence boundaries, and did occur to a minor extent in their data. But as for the palm-up sign, Hansen and Heßmann suggest that it would be most peculiar to find head nods marking the boundaries of every sentence in normal discourse.

Furthermore, eye blinks were found to co-occur rather often with sentence boundaries. For blinks, there seems to be some consistency in co-occurrence. At the moment, blinks do not seem to be conclusive for sentence identification, but suggestions for the occurrence of combinations of blinks with other cues have been provided in the past, for example by Wilbur in 1994, and also already in 1978 in an early study of nonmanuals by Baker & Padden. “Like breathing in spoken languages, a physiologically necessary brief closing of the eye may be expected to occur where it is least intrusive” (p.160). The further indicated that eye blinks in ASL combine with eye gaze related features to indicate syntactic boundaries.

In a large-scale study on sentence boundaries, Nicodemus (2009) described which prosodic cues produced by ASL interpreters were perceived by native deaf participants. Fifty deaf native signers identified sentence boundaries in a video of an interpreted lecture. Twenty-one prosodic markers were independently scored and grouped into one of the following four articulatory categories:

1. **Hands**: Held handshape, Hand clasp, Fingers wiggling, Hands drop, Signing space.
2. **Head and neck**: Head position (tilt (front and back), Head position turn (left and right), Head position: Tilt (left and right), Head movement: Nod, Head movement: Shake, Head movement (side to side), Neck tension.
3. **Eyes, nose, and mouth**: Eyebrows, Eye gaze, Eye aperture, Nose, Cheeks.

The most frequent markers were examined in each of these four categories. The results showed that in the category ‘*Hands*’ the most frequent marker is hand clasp; the second most frequent is held handshape. In the category ‘*Head and neck*’, the most frequent marker is head tilt; the second most frequent is head turn (L/R). In the category ‘*Eyes, nose, and mouth*’, the most frequent marker is eye aperture; the second most frequent is eyebrows. Finally, in the category ‘*Body*’, the most frequent marker is body lean; second most frequent is shoulders.

The cues involving larger articulators were the most frequent at boundary points (such as hand clasps and body leans). Markers of ongoing movements were used less frequently or in co-occurrence with a held marker. In addition to frequency, duration of the prosodic cues, the number of markers at each identified boundary, and the timing of the markers in relation to a target cue, which was the hand clasp. The longest duration was found for body lean and the shortest duration was found for eye aperture – which is what one would expect given the difference in mass between the articulators eye lid(s) and whole upper body. The maximum number of co-occurring cues of one sentence boundary was seven. Nevertheless, for 31% of the cues, a sequential timing pattern (occurring completely before or after a target cue) was found. The specific combinations of the occurring cues at the boundary points were not
analyzed in detail, although the overall simultaneous use of (smaller) articulators was established for most of the cues (see also Nicodemus, 2006).

One of the limitations described by Nicodemus is the use of the term ‘sentence’ in the instructions, which may not have represented the types of boundaries that were identified by the deaf participants. However, for practical reasons, this term was used to explain the task of segmenting. Which boundaries refer to real sentences remains an open question. As we indicated at the end of section 1, this is not a problem if the perspective is that of automatic

One of the further questions based on Nicodemus’ extensive work is: Can we perceive the larger cues better simply because they are perceptually larger, or are they also most frequent if we would use video analysis to examine the occurrence of cues at boundary points? In other words, are the cues that lead to the identification of a sentence boundary driven by the perceptual needs of the viewer and are people therefore incapable to segment based on smaller cues during segmenting? Related is the question whether the most frequent cues are also the most successful cues for segmentation (or do these cues at different locations in many occasions as well, which make them less successful for boundary identification). In addition, do interpreters perhaps produce the different cues differently when compared to deaf signers?

Hochgesang (2009) similarly studied sentence identification in American Sign Language. Twenty-one deaf native/early sign language users from Gallaudet University looked at three clips of narrative. The participants were divided into three groups, receiving different instructions. Seven people were asked to identify sentences, seven were asked to identify where the periods should be, and seven were asked to identify where the narrative could be divided. The first time they saw the video they were instructed to look only without doing anything. On the second viewing, they segmented the data by reporting the time code of the video where they saw the end. The participants were instructed that they could change their answers if they wished. The results showed that the type of question asked to identify the boundaries of sentences (or sentence-like units, as the author refers to it) does not have much effect. Hochgesang states in conclusion that the exact type of unit that was segmented is not quite clear. Transcription of sign language videos can be done at the level of intonation unit, utterance, idea unit, clause, sentence, Intonational Phrase, and possibly at yet other levels. In her study, she did not examine (yet) the content of the chunks that were identified by the deaf participants.

Equative sentences formed the subject of an investigation of Finnish Sign Language (FinSL) by Jantunen (2007). Equative sentences are nominal structures that are often applied for identification, such as introduction, defining and naming. In those equative sentences, Jantunen also studied the non-manual behaviors, including prosodic features such as eye blinks, eye gaze, eye brow movements, and movements of body and head position. Jantunen found that the non-manual behaviors in the different types of equative sentences showed substantial uniform occurrence of features. Similar to studies of blinks in American Sign Language (Wilbur, 1994, 2000) and Hong Kong Sign Language (Sze, 2004) Jantunen showed that in FinSL too, blinks were present at sentence boundaries. However, blinks did not always occur at that location, and moreover, blinks did also occur at sentence-internal phrases and in places other than sentence or phrase boundaries, for example within longer fingerspelled sequences. Sentence initial noun phrases showed combinations of widened/squinted eyes and raised/frowned eye brows. At the end of sentences, a head tilt was often observed. In general, alterations of head postures and also of body postures seemed to mark phrase or sentence boundaries, cf. the findings on ASL and ISL reported above.

According to Brentari (2007), native signers and non-signers do sometimes differ in their segmenting strategies. In a segmentation study, signers and non-signers were asked to mark the edges of Intonational Phrases in passages of ASL which contained pairs of identical
Signs, either with an IP break between the identical signs or not. Native signers were more accurate at detecting the IP boundaries than non-signers.

In their work on identifying clauses in signed languages, Johnston & Schembri (2006) found that signals such as pauses, blinks, changes in eye gaze, changes in brow position, changes in head position, and so forth, do not always systematically occur at boundaries. This suggests that any potential boundary cues are not completely grammaticalized, and most of these cues have a pragmatic function instead. As a result, seeing sentences in sign would present a challenge for linguistic analysis (Johnston & Schembri, 2006).

In a paper by Kingston (1999) concerning the prediction of phonological analyses through experimental investigations of phonetic behavior, three issues in the analysis of prosody of signed languages (by laboratory phonologists) were described: the internal structure of the signed syllable, realization of lexical and phrasal prominence, and the marking of edges. Given the focus on sentence boundaries in the present study, findings on the marking of edges will shortly be discussed. The two topics discussed by Kingston are final lengthening and external sandhi. Similar to spoken languages, final lengthening marks the end of phrases. Work by Wilbur and Zelaznik (1997) was described by Kingston (1999), which showed that signs were longer in final position than in other positions, but the velocity or displacement were not larger in final positions. However, this was different for phrase prominent signs, which were not longer in final positions than in other positions, however, phrase prominent signs do seem to have larger peak velocity and displacement. With respect to sandhi, it has been known that the pronunciation of the end of spoken words may influence the pronunciation of the edge of the next word, either as assimilation in case of categorical influence of pronunciation, or by overlap in case of gradient influence of pronunciation. The presence or absence of assimilation processes can also be informative about the presence of prosodic boundaries. This has not explicitly been taken up in subsequent research on sign prosody as far as we know. However, the studies about the spreading of the non-dominant hand (e.g. Sandler, 1999a; Brentari & Crossley 2002) can be related to assimilation, as spreading is argued to stop at a prosodic boundary (typically a that between two Phonological Phrases) and not go beyond it, even if the subsequent signs would allow it in being one-handed, for example.

Herrmann (2009) performed a study on prosody in German Sign Language (DGS), based on eight native DGS signers who were recorded for two hours each. Multiple prosodic cues were analyzed, referring to rhythm, prominence, or intonation. For rhythm, the following cues were analyzed: pauses, holds/frozen signs, lengthening, eye blinks, signing rate, head nods, reduplication, gestures. For prominence, the following cues were analyzed: head movement, eye brow movement, eye aperture, tense signing, lengthening and enlarging of sign. For intonation, eye brow movement, eye aperture, eye gaze, frown, facial expression, mouth gesture, and head movement were studied. Some cues are spread across multiple syllables, and function as domain markers. Domain markers that change at phrase boundaries include facial movement, head movement, and body movement (Herrmann, 2009). Edge markers are observed at the prosodic phrase boundaries, for example, eye blinks, head nods, pauses, repetition of signs, holds, and final lengthening. Around a third of the blinks did not have a prosodic function according to Herrmann’s analysis. At 78.8% of the Intonational Phrase boundaries, a blink was observed. In 94.7% of the Intonational Phrase boundaries, either blinks or other cues were observed. As in the previous studies discussed, there appears to be a complex interplay between prosodic markers as opposed to a one-to-one form and function relationship.
2.3.3 Conclusion: combinations of phonetic cues to prosodic boundaries

The presentation of the rather diverse set of studies above has made clear, there is no evidence for a dominant role of one or more cues in the signaling of prosodic boundaries. Multiple cues of both a durational and a punctual nature appear to be present in various sign languages including ASL, DGS, FinSL, NGT and BSL. Many authors point to the complex relation between syntax and phonetic form, most authors agreeing that there is a phonological level of organization including a hierarchical set of prosodic domains mediating between the two, cf. the speech model of Shattuck-Hufnagel & Turk (1996) that was presented in Figure 1 and cf. the seminal work of Nespor & Sandler (1999).

In the next two sub-sections, we will briefly discuss how machine processing of speech and sign is trying to automatically recognize prosodic boundaries.

2.4 Automatic detection of sentence boundaries in spoken languages

In comparison to sentence boundary identification of textual data, the work on sentence boundary identification in spoken languages is still relatively new. Currently, sentence boundary information can be extracted from audio with reasonable reliability. Gotoh & Renals (2000) described an approach whereby sentence boundary information was extracted statistically from text and audio resources in broadcast speech transcripts. Pause duration information based on speech recognizer outputs was used to establish boundaries, in addition to the conventional language model component that can identify sentence markers to some extent. The combination of the pause duration model and the language model provides most accurate identification of boundaries.

As for text, it is important in spoken language understanding to find the location of sentence boundaries. In text, punctuation is structurally provided. This is not explicitly indicated in spoken language. Similar to Gotoh & Renals (2000), Stolcke, Shriberg, Bates, Ostendorf, Hakkani, Plauche, Tur & Lu (1998) found that combining models (in their case a combination between prosodic and language model sources, modeled by decision trees and N-grams) led to better results than use of the individual models (see also, e.g., Shriberg, Stolcke, Hakkani-Tur & Tur, 2000). For their study, Stolcke et al. (1989) examined altogether three aspects of prosody: duration (of pauses, final vowel and final rhymes, normalized both for segment durations and speaker statistics), pitch (F0 patterns preceding the boundary, across the boundary, and pitch range relative to the speaker’s baseline), and energy (signal-to-noise-ratio). These machine processing strategies clearly indicate that for spoken language, too, no single cue will ever be reliable enough to segment the stream of language production.

2.5 Automatic detection of sentence boundaries in signed languages

Although sentence boundary detection in spoken languages is relatively new compared to textual data, (sentence) boundary detection in signed languages is even more recent. Only a very limited amount of work has been done thus far. Nevertheless, it is as important as for spoken language to detect boundaries, as for neither modality something like punctuation in text can be relied on. In a multidisciplinary research project by Koskela, Laaksonen, Jantunen, Takkinen, Raino & Raike (2008), computer vision techniques for the recognition and analysis of gesture and facial expressions from video are applied to the area of sign language processing of FinSL. Existing video feature extraction techniques provide the basis for the analysis of sign language videos. Koskela et al. (2008) further applied an existing face detection algorithm to detect the eyes, mouth, and nose. To track the motion part of the signs, they applied a standard algorithm based on detecting distinctive pixel neighborhoods. The relation between motion and sign language sentence boundaries was studied in a pioneering study in the area of signed languages. Results have not yet been published.
Related to the work by Koskola et al. (2008), Jantunen, Koskela, Laaksonen & Raino (2010) described a technical method to visualize prosodic data in sign languages. Similar techniques are extensively used in spoken language research in software such as ‘Praat’ (http://www.praat.org), in which speech recordings are analyzed and presented graphically. Data on prosody in sign language were similarly represented graphically and analyzed semi-automatically from the digital video materials. In the past, some attempts were done to perform linguistic analysis of motion and other parameters when the signed videos were produced in pre-determined laboratory settings under complex motion equipments and software. One of the major advantages of the new techniques used by Jantunen et al. (2010) is that videos no longer need to be produced in laboratory settings, but complex analysis of highly variable digital videos will become possible.

Three steps are taken in the analysis by Jantunen et al. (2010): skin regions of the participants are detected; 2. Motion of the skin area is tracked; 3. Motion is represented by using statistical descriptors. The technique is not ideal as yet: skin color detection cannot distinguish between skin on the hands and the face, and its success is dependent on lighting conditions and the color of clothing and background. Moreover, the distance between the signer and the camera cannot be measured effectively, so that movement in the front-back dimension has to be reconstructed, which can be a very complex task. This might get solved in the future, for example by using multiple video cameras with overlapping views.

As recently described by Piater, Hoyoux & Du (2010), the automatic recognition of natural signing is demanding due to the presence of multiple articulators that each have to be recognized and processed (fingers, lips, facial expressions, body position, etc.), and due to technical limitations such as spatial and temporal resolution and unreliable depth cues. Similar challenges were mentioned by Crasborn, Sloetjes, Auer & Wittenburg (2006) and Jantunen et al. (2010). Two parts were analyzed in videos containing conversations and monologues in sign language. The first part of the video analyses concerned detailed face tracking, extracting facial expressions such as mouth and eye aperture and eyebrow raise. The second part of the video analyses concerned hand tracking. Further parts will be developed in order to increase the overall tracking accuracy.

Any type of numeric data can be displayed along with video recordings in recent versions of ELAN, the multimodal annotation tool. Crasborn et al. (2006) described the development of this facility in ELAN. They presented the context of the collection of kinematic recordings of finger movements. Some of the major advantages of using such data rather than the output of video processing is the high spatial and temporal resolution that can be obtained by many systems (often around 100 Hz compared to 25 Hz for PAL video). Among the possibilities for analyses based on raw position data retrieved are the calculation of velocity, acceleration and jerk, parameters that have been argued to be informative about stress (Wilbur REFS) that may also turn out to be informative about other aspects of prosody. One of the disadvantages of using kinematic equipment is the unnatural signing environment and the impossibility to analyze the growing number of video corpora on sign languages. However, as opposed to the method described by Jantunen et al. (2010), skin color does not need to be detected for this technique, and distance to the camera is no longer an issue, given that movement in all three spatial dimensions is recorded with equal accuracy.

Cutler et al. (1997) stated the following: ‘Of great value to future work for spoken languages are greater phonetic precision, consideration of cross-language variation, and a theoretical framework allowing explicit prediction (of prosody) towards processing effects’ (p. 171). The same is true for future work related to signed languages. It will be clear that the technical possibilities that have been discussed in this paragraph have not yet led to knowledge of segmentation of signed discourse into signs and sentences, but that these techniques hold great promise for the future. Knowledge of phonetic cues is slowly growing,
including knowledge of prosodic cues of sentences. Nevertheless, none of the investigated prosodic cues seem to provide a fully reliable predictor function thus far for the presence of sentence boundaries. As Sandler stated in 2006: “Neither instrumental tracking and transcribing of the prosodic system in sign language, nor experimental work on its perception and interpretation, have yet been done” (p. 265). In the next section, two possible research directions are described that can contribute to our understanding of prosodic cues at sentence boundaries, one involving experimental work on human perception and one involving instrumental tracking.

3. Suggestions for two types of empirical studies

First of all, we suggest new tests of human segmentation of signed sentences. Secondly, we sketch how the use of new tracking techniques from visual signal processing can be engaged to detect salient events in a video recording of sign language.

3.1 Study 1: New tests of human segmentation of signed sentences

We would like to suggest several new tests of human segmentation of signed sentences which all include video manipulation of various kinds. The goal remains to gain further insight in the human detection of prosodic cues. The overall idea behind the tests is to prevent the semantic processing of the signing in the video to human subjects, whether signers or non-signers, as that interferes with the pure phonetic parsing. Similar techniques have been used in studies of spoken languages whereby the spoken stream is manipulated in such a way that it becomes harder to understand the speech, while people could process prosodic features (van Bezooijen & Boves, 1986; Mettouchi, Lacheret-Dujour, Silber-Varod & Izre'el 2007). Typically, low-pass filters are applied which obfuscate the segmental content while maintaining duration and melodic properties.

A parallel signal manipulation for sign language videos could involve a strong decrease in visual quality, e.g., by blurring the visual scene or by lowering the spatial resolution (dots per inch). Furthermore, if the aim would be to examine the specific contribution of prosodic features of the head and face in particular, separately from the contribution of the body and the manual prosodic features, in an additional video manipulation only the face of the signers could be shown without showing the remainder of the body. Conversely, if the aim would be to examine the specific contribution of prosodic features of the body and the hands, only the body of the signer could be shown without showing the head. Those suggestions are expected to strongly disrupt the processing of semantic information in the discourse. Alternatively, prosodic cues can be made to stand out in a (manipulated) video in order to make the prosodic cues more noticeable. For example, the eye contours can be highlighted to make short blinks more easily detectable, or body and head contours can be highlighted (e.g., change colour when the head or body move in a certain way).

Quite a different way of hindering the semantic processing of the sign stream would be to process an unfamiliar (sign) language (e.g., Fenlon et al., 2007, for sign languages; Mettouchi et al., 2007, for spoken languages). In contrast to the processing of an unfamiliar spoken language, the processing of an unfamiliar sign language may be affected by the presence of some similarities between sign languages, especially for those signs that are semantically motivated. Fenlon et al. reported that the signers of British Sign Language in their study were able to partly understand stories in (unfamiliar) Swedish Sign Language. Signers of the familiar and unfamiliar sign language reported similar prosodic boundaries,
which may emphasize the benefit of using video manipulation to hinder the understanding of the content. Fenlon et al. also involved hearing non-signers, who may have relied on cues they knew from gestures in face-to-face communication.

Another problem in having users of a different sign language process the prosody of a particular sign language is that the prosodic systems and the phonetic cues that are used differ between the two languages. While the literature review in section 2 has mainly shown overlaps between languages in the type of cues that are used, it may still be the case that the precise timing and quality of the nonmanual articulation varies. In fact, given what we know about spoken language prosody, it is highly likely that linguistic variation between sign languages is also located in the phonetic implementation of phonological features (e.g. Gussenhoven, 2004). For that reason, we suggest that preventing access to semantic processing by image manipulation is to be preferred over the use of subjects that do not master the sign language in question – whether foreign sign language users or non-signers.

More generally, independent of the precise method of eliciting human segmentation judgments, post-hoc analyses of prosodic cues at high-likelihood segmentation regions would be useful to examine the co-occurrence of various combinations of cues (the end and start of domain markers in combination with boundary markers), as past empirical studies have shown that none of the individual prosodic cues provide sufficient predicative power for the occurrence of a sentence boundary. Several people have emphasised the presence of co-occurring cues at boundaries, however, specific combination of prosodic cues have not been specified thus far.

The video manipulations and the general suggestion for post-hoc analyses of co-occurring cues may provide new insights into the use of prosodic cues at sentence boundaries.

3.2 Study 2: Experimental tests of human vs. machine segmentation

In sign language research concerning prosody, video analysis and the use of cyber gloves could also be applied to extract useful information on prosodic cues. Techniques such as described by Piater et al. (2010) and Jantunen et al. (2010) in combination with new tools such as described by Crasborn et al. (2006) are very promising with respect to the analysis of prosodic cues in non-restricted natural continuous sign language. One of the main advantages is that large corpora can be processed this way. This in turn enables statistical patterns among the co-occurrence of different types of cues can then be calculated. These techniques would allow refinements of the recent human perception studies in which the intuitions of signers and non-signers were analyzed concerning boundaries in sign language discourse. It may prove useful to use data deriving from video analysis and cyber gloves techniques to examine the actual co-occurring of specific combinations of prosodic cues that appear to be most predictive for sentence boundaries cues. We should emphasize, however, that the discovery of statistical patterns in large data sets does not in itself constitute a linguistic analysis of the structures in question. They should be considered as tools for linguistic analyses, just as quantitative analyses are used in spoken language phonetic research. It is the linguistic model that should generate the hypotheses to test. For automatic recognition and translation of signed languages, however, there are more direct advantages of computer processing of phonetic cues.

In Figure 2, an example is shown of feature extraction of the face, developed by Piater et al. (2010, see also Dreuw et al., 2010). For each of the four video images, three drawings are presented of the fitted model. From top to bottom, the following models are presented: a full model instance, a meshed shape, and a plotted shape. In addition, three vertical lines are present in the image, quantifying several nonmanual cues. From left to right, the following features are presented: left eye aperture, mouth aperture, and right eye aperture. The three
axes on the face represent information on the orientation of the face: with the origin at the tip of the nose, the red line is the X axis, the green line the Y axis, and the blue line is the Z-axis.

Figure 2. Visual representation of automatically detected properties of nonmanual visual cues. Images courtesy of Justus Piater and Thomas Huyoux, Université de Liège, Belgium.

Shattuck-Hufnagel and Turk (1996) emphasized that in studies of prosody on spoken sentences, acoustic measurements should ideally be complemented by perceptual measurements of the prosodic structure, since it is difficult to detect prosody based on the signal alone. This is particularly true in the case of sign languages as well, given that retrieving prosodic information of the sign signal is in a very early stage. Currently, some prosodic signals can be detected more easily than others by software, depending on several (sometimes restricting) factors, such as exceptionally large or small movements, the visual similarity between the skin of the hands and of the face when manual signs occur in front of or nearby the face, and the general quality of video materials, only to name a few.

Regardless of these challenges in video analyses, video analyses and intuitive judgments can be mutually highly informative for the increase of our knowledge on sentence boundaries. Video analyses can provide exceedingly detailed data on each of the possible cues. However, intuitive judgments can provide information of the actual presence of a prosodic boundary in sign language. Without native informants to confirm the feature extraction data during the initial phase of analysis of (combination of) features deriving from video analyses, it would be impossible to judge whether the data in fact pointed towards boundaries or not. Beyond the current technical restrictions in video analysis, the equivalent of the ‘acoustic’ measurements based on automatic measures should therefore be combined with the perceptual transcription of the prosodic boundaries. In an ideal situation, the perceptual measures and transcriptions should be elicited from (native or near-native) signers who can provide intuitive judgments on their own language.

As was already noted above, additional syntactic analyses would subsequently be necessary to provide more information on the actual domains that are identified. Equally, the identification of boundaries by intuitive signers can be analyzed into much more depth if not only manual annotations of (co-)occurring cues were provided for the identified boundaries,
but also information based on the feature extraction data that cannot be inferred in detail from the naked eye, such as temporal details (duration, velocity, and acceleration).

4. Discussion

How can knowledge of the prosody of signed languages be employed for language technology, such as the automatic translation from signed to spoken language? As long we have not captured reliable predictors (specific combinations of prosodic cues) of sentence boundaries and a reliable method to capture these prosodic cues, at least three challenges will be present. First of all, a time-consuming manual procedure will remain to be required to annotate each sentence boundary, based on native signer’s intuitive judgments. Secondly, automatic translation will not remain very difficult until prosodic sentence boundary cues can be detected from video analyses. Finally, as long as sentence boundary annotations need to be added manually, a large range of interpersonal variation will occur and some prosodic cues might not be observed in sufficient detail. This implies that increased knowledge of prosody of sign languages and prosody related to sentence boundaries in particular can lead to better predictions of the presence of boundaries.

Once these predictive cues can be captured automatically and reliably by means of video analysis (if possible, combined with the capturing of lexical items with specific shapes that often occur at sentence boundaries), sentence boundaries can be automatically generated at the correct boundary locations. This can facilitate the analysis of larger data sets than would have been possible by phonetic transcription. Moreover, the captured chunks can theoretically be translated automatically. The reason that ‘theoretically’ was used is that sentence translation can only commence from the moment a sufficient number of signs in the sentence can be automatically recognized by machine recognition. Many challenges are still to be overcome in this area, given our restricted understanding of sign languages (Ormel et al., 2010).
Ormel & Crasborn (subm.), The prosodic correlates of sentences in signed languages and their relation to language technology: a literature review. Sign Language Studies.

References


Crasborn, O. (2007). How to recognize a sentence when you see one. Sign Language & Linguistics, 10, 103-111.


