4.1 Introduction

Prompted by the desire to test claims about procedural modularity of the language and vision systems (Fodor 1983), psycholinguistic research has examined the interplay of language comprehension with visual perception. This interplay is rapid, incremental, and closely temporally coordinated. It further appears to be bidirectional such that language guides visual attention and attended non-linguistic visual information in turn affects comprehension (henceforth ‘visually situated’ language comprehension). Information from a visual context can crucially inform syntactic disambiguation within a few hundred milliseconds, a finding that suggests core comprehension processes are not procedurally modular (Spivey et al. 2002; Tanenhaus et al. 1995). These and other findings have shaped accounts of visually situated language processing in which language guides (visual) attention and attended non-linguistic information can rapidly influence comprehension processes (e.g., Altmann and Kamide 2007; Knoeferle and Crocker 2006, 2007).

However, most eye-tracking evidence on this temporally coordinated interplay comes from studies with healthy university students (approximately 18–31 years of age), thus focusing on a short interval of adult life and on a small population segment. To ensure the generality of language comprehension accounts, we could test them more broadly. This includes tests across the lifespan (e.g., in infancy, childhood, young adulthood, and older age, e.g., Carminati and Knoeferle 2013; Trueswell et al. 1999; Zhang and Knoeferle 2012), on diverse social population segments (e.g., working class people, academics, illiterates, or dialect speakers, e.g.)
Huettig et al. 2011, 2012), and for different levels of cognitive function (e.g., good vs. poor comprehenders or people with high vs. low working memory, e.g. Nation et al. 2003; Knoeferle et al. 2011b). This would bring us closer towards including a model of the comprehender (and ultimately also, of the speaker) into visually situated language comprehension accounts and permit us to refine their predictions.

With regard to insights across the lifespan, a number of studies examining young infants and children could inform us about this interplay. However, relatively few studies have undertaken a “direct” (i.e., within-item) comparison of visually situated child and adult language comprehension. Strictly speaking, this is necessary to exclude the possibility that between-group differences in the time course or manner of processing result from stimulus variation. The results from one influential within-item comparison in 5-year olds and adults suggested marked differences between these two groups (Trueswell et al. 1999; see also Snedeker and Trueswell 2004). Overall, the children relied more on their linguistic knowledge for syntactic structuring than the adults. By contrast, children failed to exhibit an adult-like reliance on the visual referential context for syntactic structuring and disambiguation.

These results suggest that the non-linguistic referential context is of limited importance for at least some processes in child language comprehension (e.g., real-time syntactic disambiguation of local structural ambiguity). The rapid, temporally coordinated interplay of language comprehension and visual attention in adults also seems to emerge gradually in infants. While 24-month-olds did not rapidly relate visual referential contrast to an unfolding utterance, 36-month-olds resembled adults in that they shifted their visual attention more quickly to a target picture upon hearing “blue…” (“blue car”) when the context contained only one blue object (a blue car) than when it contained two blue objects (Fernald et al. 2010).

The insights into the limited role of the visual context for real-time language comprehension seem to contrast with its role in language learning. Reference from nouns to objects across different learning instances, for instance, plays an influential role for language learning (e.g., Smith and Yu 2008; Yu and Smith 2010). Other cues such as object sequences also seem to be important for guiding infants’ visual attention to, and their learning about, objects (e.g., Wu et al. 2011). Infants are further sensitive to actions and to goal-related information before they complete the first year of their life (e.g., Woodward and Sommerville 2000). In addition, non-linguistic behaviours such as joint attention episodes seem positively related to young infants’ vocabulary size (e.g., Tomasello and Farrar 1986; Tomasello and Todd 1983). Thus, subtle cues in the non-linguistic visual context appear to play an important role for learning about language and objects.

It is possible that these differences in comprehension compared with learning result from different grain sizes in the time course of these two processes. For language comprehension, children must orient towards a visual stimulus within a few hundred milliseconds. Comprehension studies often do not repeat individual stimuli, and the child has a window of no more than a few hundred milliseconds to relate the visual context to language. Language learning success, by contrast, is often measured offline. And in those studies that measured eye gaze, stimuli were repeated and rapid orienting to a mentioned object could thus have emerged over time.
Another possibility is that non-linguistic visual context plays a role only for some cognitive processes (e.g., for establishing reference but not for syntactic disambiguation). The picture may further be complicated by variation as to when in the child’s development visual cues affect specific comprehension processes (see Trueswell et al. 1999, pp. 122ff. for discussion of children’s reliance on linguistic knowledge and their insensitivity to the visual referential context).

4.2 Summary

In sum, the role of the non-linguistic context for incremental language comprehension across the lifespan remains unclear. And yet, this is an important point for refining predictions of visually situated language comprehension accounts with a model of the comprehender. This chapter accordingly argues in favour of examining and modelling the realtime interplay of language comprehension and visual attention more broadly. I will illustrate this argument by comparing visually situated language comprehension in children and in adults. Visual context could serve merely as a backdrop to children’s language-mediated visual attention. Alternatively, it could play a more active role, not just in language learning but also in comprehension. A first section motivates the active role of the visual context for comprehension by appealing to insights from learning about language, objects, and events (sect. 4.3). A subsequent section reviews insights into the interplay of language comprehension with visual attention, as well as into visual context effects on language comprehension (sect. 4.4). Based on this review, I will argue that (attention to objects and events in) the non-linguistic visual context plays a fundamental role not only for learning about language and objects but also for child language comprehension.

4.3 Learning About Objects and About Language in Visual Context

For a temporally coordinated interplay between language comprehension, visual attention, and visual context effects, children must be able to rapidly map nouns onto objects, and actions onto verbs. One pre-requisite for this is that children can identify words and objects from the stream of auditory and visual events. Indeed, infants can segment words from the speech stream through, for instance, statistical regularities. 8-month-olds listened to a stream of artificial speech containing “words” (e.g., “tilado”) which could be detected based on transitional probabilities of syllable pairs (Saffran et al. 1996). Within words, the transitional probability of

---

1The transitional probability (X|Y) of a syllable pair XY is given by the frequency of the pair XY divided by the frequency of X.
syllable pairs was set to 1.0 (e.g., for “tilado”, “la” always followed “ti”), and across word boundaries, it was set to 0.3. Infants’ listening times to words compared with non-words were inferred from how long they fixated a blinking light on a side-wall of the test booth. Crucially, they fixated the blinking light longer when listening to word than non-word stimuli, suggesting word recognition.

Infants’ sound processing is further finely attuned to the rhythmic structure of their language, and it has been suggested that this sensitivity plays a role for word segmentation. Höhle et al. (2009) played CVCV sequences (e.g., “gaba”) with either trochaic or iambic stress to German 6-month-olds. Both stress patterns had a lengthened second syllable and a higher-pitched first syllable. They differed, however, in the degree to which they exhibited these two patterns. For iambic (compared with trochaic) metre, the increased duration of the second over the first syllable was more pronounced; and for trochaic (compared with iambic) metre, the pitch of the first syllable was higher. The German 6-month-olds preferred to fixate a blinking light on the side-wall of the test booth longer for the trochaic than iambic stimuli, a preferential looking behaviour that generalised neither to German 4-month-old infants (but see Friederici et al. 2007) nor to French 6 month-olds. It seems then that the specific rhythmic pattern of German modulated the German infants’ use of stress cues, a factor that can influence word segmentation (see also Jusczyk et al. 1993, 1999).

Infants can also learn verbs from a very early age, a process that benefits from the sentential syntactic structure (e.g., Fisher et al. 1994; Gleitman 1990; Naigles 1990). Children inspected scenes (e.g., a rabbit feeding an elephant with a spoon) and responded to questions about the meaning of novel verbs (e.g., “zorking”) in one of three syntactic contexts (transitive, intransitive, and uninformative verb contexts, Fisher et al. 1994). The 2-year-olds interpreted the novel verb more often as reflecting feeding when the syntax supported feeding (i.e., “The bunny is …ing the elephant.”) than when it supported an intransitive event (“The elephant is …ing.”). Thus, children can assign meaning to a novel verb based on its syntactic context.

There has been discussion as to when abstract syntactic knowledge emerges. While some scientists advocate a late emergence and an early reliance on lexical learning (e.g., Tomasello 2000; Tomasello and Brooks 1998), others argue that abstract syntactic knowledge (e.g., of verb transitivity) is key for early verb learning (e.g., Gleitman 1990; Fisher 1996, 2002; Pinker 1994). Either way, children begin to learn verbs such as “dance” and “open” by around 10 months of age (e.g., Fenson et al. 1994), and verb-related syntactic knowledge from approximately 2 years of age, if not before (Fisher 2002; Golinkoff and Hirsh-Pasek 2008).

Learning of statistical regularities appears to be domain-general and extends to object sequences (Kirkham et al. 2002). After infants (2-month-olds, 5-month-olds, and 8-month-olds) had inspected object sequences with different transitional probabilities (within-pair: 1.0; between-pair: 0.3), they looked longer at novel than highly familiar object sequences (i.e., with 1.0 transitional probability). Infants’ age had no reliable effect on their novelty preference. However, another study observed longer inspection of novel than known circle sequences in 11-month but not in 8-month-olds (Kirkham et al. 2007, Experiment 1).
Results on the learning of object regularities generalised even to within-object feature-changes (Wu et al. 2011). In a first experiment, Wu et al. familiarised 9-months-olds with how objects split into parts by showing them two instances of an object’s splitting behaviour. During testing, the infants saw either a consistent or an inconsistent split of a familiar object. They gazed significantly longer at objects splitting in an inconsistent than in a familiar way. Beyond information from the objects themselves, the infants used a speaker’s eye-gaze for learning object features. Learning improved and children preferred the object with the inconsistent (vs. consistent) split, when the speaker had inspected it (but not when no speaker was visible, Wu et al. 2011). This generalised both when only one object was present and when two objects (a target and a distractor) competed for attention.

Just as infants can learn object-based regularities, they can also learn about actions and events (Sharon and Wynn 1998; Wynn 1996). In a habituation paradigm, 6-months-olds inspected a puppet jumping either two or three times. At test, they were able to discriminate these two events from a sequence of two- and three-jump events. By 14–17-months of age, they were also able to discriminate motion from path events (Pulverman et al. 2008) and 18-month-olds anticipated the goal of incomplete actions (Meltzoff 1995). Six-to-nine-month-olds habituated to actions in which a human agent (vs. an inanimate rod) reached for one of two presented toys. During testing, an experimenter switched the location of the two toy objects and the actor either executed the action with a different arm trajectory or grasped the other toy. Infants inspected a changed toy longer than a changed arm trajectory and this gaze pattern emerged for animate agents only (Woodward 1998).

Interestingly, infants can also use familiar actions to interpret novel and potentially ambiguous actions. Twelve-month-olds inspected a hand touching the lid of a clear box containing a toy animal. That action could be interpreted as directed at either the box or the toy animal (Woodward and Sommerville 2000). In one condition, the actor touched the lid but did not open the box. In another condition, the actor touched the box, opened it, and grabbed the toy. During testing, the actor touched either the old box (containing a new toy), or a new box. Infants inspected the new box for the same time as the old box if the actor had merely touched the old box during habituation. However, if he had instead taken the toy, they inspected the old box longer than the new box. Grabbing a toy appeared to influence infants’ visual attention when observing a novel ambiguous action. This result failed to replicate when the toy was outside of the box (i.e., when touching the lid of the box was no longer related to taking the toy). Simply seeing the actor touch the lid of the box with the toy nearby did not affect the interpretation of the ambiguous action.

Overall, children are thus exquisitely sensitive to regularities and structure in the linguistic and non-linguistic input. If they were unable to identify both words and objects, they could not relate these two information sources in real-time. With this ingredient, however, infants could learn to rapidly relate words to objects, and to exploit this link during real-time language comprehension.

Just as children can recruit regularities in the linguistic and non-linguistic visual context from individual learning instances, they can also learn word-object correspondences from cross-situational occurrences. In one recent study, 12-month and
14-month old infants inspected two objects per trial and heard two names. Within a trial, the mapping from names to objects was ambiguous (Smith and Yu 2008) but across trials reference was unambiguous. An infant would see two novel objects during training and hear “bosa” and “gasser”. On a later trial, the infant would see the referent of “bosa” together with a new object and hear “bosa” and “colat”. Based on cross-trial co-occurrences of an object and a name (one object always occurred together with the word “bosa”), the infant could deduce the reference for “bosa”. During testing, the authors monitored infants’ attention to two pictures in response to object names. The infants inspected the named object longer than the other object. Despite referential ambiguity for individual trials, the infants related the name of an object to its appropriate referent based on cross-trial statistics.

It appears then that infants are not only finely tuned to regularities in their linguistic and non-linguistic environment but that they can also relate at least a small number of novel nouns to novel objects through cross-situational statistics. These findings alone, however, would not necessarily support the view that visual context can play an important role for child language comprehension. During language comprehension, words are processed in a rapid sequence. Relating them in real-time to objects in the visual context requires an attentional system that responds rapidly to spoken words. Adults inspect relevant visual context information within a few hundred milliseconds during language comprehension, and this relationship holds even when scenes are cluttered and objects are mentioned in rapid succession (Andersson et al. 2011). In addition, attended information in the visual context rapidly affects their comprehension and visual attention (e.g., Chambers et al. 2004; Knoeferle et al. 2005, 2008; Tanenhaus et al. 1995). To predict that visual context plays an important role for child language comprehension, we would want to see some evidence that a similarly close temporal coordination between seeing an object, hearing its name, and inspecting it, also contributes to successful language learning.

Recent evidence for this view comes from a study on cross-situational word-referent mapping. Yu and Smith (2010) recorded the eye movements of 14-month-olds in a statistical learning task and analysed the relationship between individual differences in eye-gaze pattern and learning success. During learning, thirty slides showed pictures of two novel objects that were named (e.g., “gasser”, “bosa”). Word-object reference was ambiguous for a given slide, and correct word-object pairs occurred ten times during learning. Results from a preferential-looking task showed that despite within-trial ambiguity, most infants learned more than half of the words through cross-trial statistics.

However, some infants learned hardly any word-referent mappings (weak learners) while others seemed to master five out of the six words (strong learners). Strong learners switched less often between the two objects than weak learners and made longer fixations to the attended object during training. These differences were most pronounced in the middle of the learning phase. Weak compared with strong learners had higher entropy of eye movements per trial (i.e., more attention switches between the objects within a trial and more evenly-distributed fixation times across the two objects). They crucially also had a higher entropy time-locked to word presentation. Thus, success in cross-situational learning of word-object mappings seems to be related to how the infants inspect an object during learning.
In summary, children are sensitive to subtle statistical regularities in their visual environment. They can exploit information from both referenced objects and speakers for language learning. Importantly, strong learners seemed to exhibit a temporal coordination between hearing an object name and inspecting the named object (Yu and Smith 2010). Temporal synchrony was also important for word learning in 7- and 8-month-olds (Gogate 2010, Experiment 1). An actor pronounced syllables such as “gah” and “tah” either together with moving an object (downward, laterally, or forward) or asynchronously with object motion (e.g., in the pauses between moving the objects). 8- but not 7-month-olds learned sound-object mappings in the synchronous but not in the asynchronous presentation condition (see also Gogate et al. 2006).

Much of the reviewed evidence comes from learning about nouns and objects. As discussed above, research on how infants acquire the meaning of novel verbs highlights how sentential syntactic structure contributes to verb learning (e.g., Fisher et al. 1994; Gleitman 1990; Naigles 1990). Learners also appear to be most successful in learning nouns when both the referential visual context and language contain corresponding systematic pattern (e.g., Moeser and Olson 1974 and references therein). Indeed, two key pre-requisites for enabling non-linguistic visual context effects during child language comprehension are likely children’s sensitivity to fine-grained regularities in the linguistic and non-linguistic input, as well as the role that (the timing of) attention (to object naming) plays for successful noun learning.

4.4 Visually Situated Language Comprehension in Young Infants and Adults

Based on these insights from language learning, sect. 4.4 argues that the non-linguistic visual context and its interaction with linguistic cues can play an important role for child language comprehension.

4.4.1 How Language Comprehension Guides Visual Attention in Children and Adults

To examine the real-time interplay between language comprehension and visual attention in infants, several studies have monitored infants’ eye movements in picture contexts during spoken language comprehension. Six-to-nine-month-olds rapidly shifted attention to a named object (e.g., “Look at the banana!”) in a display showing a picture of a banana and of a head (Bergelson and Swingley 2012). And when watching videos of their parents, 6-month-olds looked more often at the correct parent upon hearing “mommy” or “daddy”, a behaviour that did not emerge with videos of unfamiliar adults (Tincoff and Jusczyk 1999). Six-month-olds were even able to visually orient towards adult body parts upon hearing them named (hands compared with feet, Tincoff and Jusczyk 2012). Interestingly, 15-month-olds rapidly oriented
their eye-gaze to a familiar referent (e.g., a dog) but they did so only after they had heard the entire name (e.g., “doggie”, Fernald et al. 1998). By contrast, 24-month-olds—similar to adults—began to shift their eye-gaze to a referent before they had heard its entire name. By the age of two, children thus resembled adults in that even incomplete words guided their visual attention to objects. When adults listened to a word such as “beetle”, they inspected the picture of a beetle more often than unrelated targets from around 200 ms after word onset (e.g., Allopenna et al. 1998; see also Dahan 2010; Tanenhaus et al. 1995).

The rapid interplay of spoken comprehension and information from visual context is also evident in event-related brain potentials (ERPs): Friedrich and Friederici (2004) compared semantic processing in 19-month-olds and adults. A picture of an object accompanied by an incongruous word (vs. the object name) elicited more negative mean amplitude ERPs at right hemisphere and midline sites (from approximately 300 ms), as well as locally over temporal sites (from 100 to 250 ms) in the adults. The continued greater negativity for incongruous relative to congruous stimuli resembled the N400 effect observed for semantically incongruous relative to congruous words in sentence contexts (Kutas and Hillyard 1980, 1984). For the 19-month-olds, incongruous words triggered a broadly distributed negativity compared with congruous words which overall resembled the N400-like effect observed in the adults. The emergence of the same component in both of these groups highlights similarity in their semantic interpretation mechanisms. However, the infant N400 effect occurred 400 ms later, lasted somewhat longer, and had a different topography than the adult N400. Children are thus subtly slower in noun-object mapping, and the topography of the N400 suggests some differences in their interpretation mechanisms relative to the adults.

Children’s visual attention seems to be guided rapidly also by their grammatical knowledge. Spanish-learning children (aged 34–42 months) listened to sentences such as (Encuentra la pelota, ‘Find the ball’), while inspecting related objects in the visual context. Upon hearing ‘the ball’, they inspected the target (a ball) earlier when the object names differed in grammatical gender (“la pelota”, “ball [feminine]”; “el zapato”, ‘shoe [masculine]’), than when they had the same gender (“la pelota”, “ball [feminine]”; “la galleta”, ‘cookie [feminine]’, Lew-Williams and Fernald 2010; see also Van Heugthen and Shi 2009 for similar evidence in 2-year-olds). Children’s rapid use of gender marking in finding an object resembles that of adults. When adult participants had to click on one out of four objects, and two of these had similar-sounding name onsets (‘bouteille’ and ‘bouton’), they inspected the objects with similar names more often than the phonologically-unrelated objects for a brief time interval after word onset. This effect disappeared when a gender-marked article differentiated between the referents with similar name onsets (‘le bouton’ vs. ‘la bouteille’, Dahan et al. 2000).

Just as children can rapidly relate nouns to objects, they can also predict an object in the visual context based on verb information. When 10–11-year-olds inspected clipart scenes in which a verb (e.g., “eat”) restricted the domain of reference to just one object (a cake), they began to anticipate that object before its mention, and more often than when the verb was non-restrictive (e.g., “choose”,

---

[64 P. Knoeferle]
Visually Situated Language Comprehension in Children …

This was true independent of whether the children were skilled or less skilled comprehenders. However, compared with highly skilled comprehenders, the less skilled ones made more fixations and spent less time inspecting the target. Adults’ time course of anticipating an edible object for “eat” relative to “move” resembles that of 10–11-year-olds (Altmann and Kamide 1999). Crucially, this predictive capacity is present from an early age. When 2-year-olds listened to sentences such as “The boy eats a big cake”, they began to inspect a depicted cake soon after they had heard a semantically restrictive verb (and before “cake” was mentioned, Mani and Huettig 2012).

Even linguistic cues to abstract temporal aspects of a context rapidly influenced 3–5-year-olds’ visual attention (Zhou et al. 2014). The children and adults listened to sentences in Mandarin Chinese about an old lady planting a flower. The verb ‘plant’ appeared either together with the particle “le”, indicating the perfective aspect of the event (e.g., “zhong-le”, ‘has planted’), or with the particle “zhe”, indicating that the event was ongoing (e.g., “zhong-zhe”, ‘is planting’). During comprehension, the participants inspected a computer screen showing two pictures. One of them depicted the completed version of the event (the flower was in the soil and the old lady was walking away) while the other depicted the old lady as she was planting the flower. Both, the children (3-, 4-, and 5-year-olds) and the adults began to inspect the completed event picture more often approximately 200–400 ms after the onset of “le” than “zhe”, and they began to inspect the ongoing event more often 200–400 ms after the onset of “zhe” compared with “le”. This gaze pattern did not vary with age, suggesting that aspectual markers can guide both young children and adults’ visual attention to the temporal structure of depicted events.

Children moreover resemble adults in how lexical cues guide their attention during syntactic disambiguation but they can experience delay in exploiting prosodic cues. In Snedeker and Yuan (2008), 4–6-year-olds inspected four objects (e.g., a candle, a feather, a frog with a feather, and a leopard with a candle), and were instructed that they “can feel the frog with the feather”. Sentences in Experiment (1) had either instrument prosody (with an intonational phrase break after the object noun) or modifier prosody (with an intonational phrase break after the verb). Looks to, and actions with, the instrument were taken to reflect interpretation of the ambiguous prepositional phrase “with the feather” as an instrument. The children inspected the target instrument (the feather) more often from around 700 ms after the onset of the prepositional phrase (“with the feather”), and used the feather more often for action execution when they heard the utterance in instrument rather than modifier prosody.

In a second related study, the authors manipulated verb bias in addition to prosody. The verb either biased towards a modifier interpretation (e.g., “choose the cow with the stick”), was neutral (“feel”) or biased towards the instrument

---

The prosodic manipulation was similar to Experiment 1; prosodic analyses ascertained accent differences between the instrument (mostly H* on the verb and noun and a pitch accent on “with”) and the modifier intonation (mostly an L + H* accent on the verb, a clear H* on the noun, and no pitch accent on “with”).
interpretation (“tickle the pig with the feather”). Children, much like the adults, used the verb bias within 200 ms after the onset of “feather” for structural disambiguation, but were delayed in their visual response to prosodic cues by approximately 500 ms relative to the adults (see also Weber et al. 2006 for relevant prosodic effects in adults).

Prosodic cues such as contrastive pitch accent is used rapidly by adults and young children alike, although children benefit from additional time in their prosodic processing. Ito et al. (2012) presented Japanese 6-year-olds with 2-utterance sequences in which the first singled out one among many depicted objects (e.g., a pink cat). The second utterance mentioned another object which was either of the same category (e.g., a green cat) or of another category (e.g., an orange monkey), and the noun phrase referring to that object either carried a pitch accent on the pre-nominal colour adjective or not. When adults and children processed these kinds of stimuli, the contrastive pitch accent prompted faster looks to the green cat when “green” was accentuated. Adults (but not children) showed garden-pathing when the utterance sequences were not contrastive (i.e., the second utterance mentioned the orange monkey). In this case the adults temporarily fixated the target signalled by the contrastive pitch (the orange cat) while children showed no such effect. When they were given more time between the two utterances, however, the older among the 6-year-olds began to show garden-pathing similar to that evident in the adults. Thus, the children seem to process these prosodic cues rapidly when they are felicitous, and when they are infelicitous, they can use them incrementally if they are awarded extra time.

When prosody served a different purpose (signalling the illocutionary status of an utterance as a statement or a question), children relied on it as rapidly as adults. Zhou et al. (2012) presented Mandarin Chinese utterances to 4–5-year-olds and to adults. The utterances were identical but intonation disambiguated them as either a statement (level intonation on the “wh”-phrase, e.g., literally ‘Xiaoming not pick what fruit’, sense: ‘Xiaoming did not pick any fruit’) or a question (rising intonation on the same phrase ‘what fruit’, sense: ‘What fruit did Xiaoming not pick?’). The visual context showed a boy with bananas, two distractor animals, and two further pieces of fruit (e.g., a pear and an orange). When the utterance was disambiguated as a question (vs. statement), participants more often inspected the fruit that the boy had not picked (the pear and the orange) and responded to the question. For statement compared with question intonation, they inspected the boy with the bananas more often and decided whether the sentence matched the depiction. Importantly, the children did not differ from the adults in how rapidly prosody affected their visual attention, suggesting similarities in children and adults’ real-time use of prosodic cues for the disambiguation of illocutionary force during comprehension.

By contrast, pronounced differences in the time course of child and adult language processing seem to emerge for pragmatic inferences. Huang and Snedeker (2009b) showed their adult participants pictures of two boys and one girl, and an experimenter distributed nine socks evenly among them (Experiment 3). Upon hearing “Point to the girl that has some of the socks”, the adults rapidly inspected
the girl that had three socks. By contrast, when “some of the” was referentially ambiguous (the boy and one girl were each given two socks and another girl received three balls), adults inspected the girl with two socks substantially later for “some” (around 1000 ms after quantifier onset) compared with “all” and compared with “two” or “three” as scalars (target inspection was above chance 200 ms after quantifier onset, Experiments 1 and 2). The authors attributed this delay to the computation of scalar implicature since gaze pattern suggested that “some” was interpreted without delay when its meaning was sufficient to disambiguate reference. Compared with the adults, 5–9-year-olds interpreted the quantifier “some” much later, after phonological disambiguation of the target noun (“socks”), suggesting that they, unlike the adults failed to compute pragmatic implicature in real-time (Huang and Snedeker 2009a).

Thus, children can relate words to objects from as early as 6 months of age. They employ their grammatical knowledge as well as semantic information related to adjectives and nouns similar to the adults from approximately 24 to 36 months of age. For verbs, adult-like visual inspection of objects associated with the verb emerges also around 24 months of age. Lexical (verb bias) information affects 4–6-year-olds’ structural disambiguation as rapidly as adults’ structural disambiguation. Children’s use of prosodic cues during comprehension, however, appears to be somewhat more variable. When intonational phrase breaks and differences in accentuation served as cues to structural disambiguation, visual attention shifts to the target objects were subtly delayed in 4–6-year-olds relative to adults. Similarities between children and adults emerged, however, in Japanese 6-year-olds’ use of contrastive pitch accent when that accent was contextually felicitous, and also in Mandarin Chinese 4–5-year-olds when prosody cued the illocutionary force of an utterance. Clear age differences emerged for complex pragmatic processes. Even much older children (5–9 years of age) failed to compute scalar implicature in realtime, a finding hinting at differences in child and adult language comprehension at the pragmatic level.

4.4.2 Non-linguistic Visual Context Effects on Comprehension and Visual Attention

In the preceding section, we reviewed how language guides visual attention. This section focuses on how visual context influences language comprehension once it has been identified as relevant by language. One key question for accounts of visually situated language comprehension is to which extent the non-linguistic visual context modulates children’s language comprehension in real-time. For adults, different visual cues are integrated within a few hundred milliseconds. Among these is information about an object’s size, colour (Sedivy et al. 1999; Huettig and Altmann 2011) and shape (Dahan and Tanenhaus 2005) but also depicted clipart events (Knoeferle et al. 2005), real-world action events (Knoeferle et al. 2011a) and action affordances (Chambers et al. 2004). For noun-object relationships, the
language learning literature supports the prediction of rapid visual context effects also in child language comprehension. For verbs and actions, the case is less clear. If the children do resemble adults in language comprehension, however, rapid visual context effects should emerge both when a noun and when a verb identifies relevant aspects of the visual context. For adults, it has even been argued that they ground their language comprehension in the immediate or recent events in preference over relying on their linguistic and world knowledge (Knoeferle and Crocker 2006, 2007). Below we discuss to which extent these characteristics of adult comprehension extend to children.

Sekerina and Trueswell (2012), for instance, failed to observe rapid visual context effects in Russian 6-year-olds. They presented the Russian children with spoken utterances (e.g., ‘Put the red butterfly in the bag’) while the children inspected a visual context depicting nine objects. The visual context either contained only one contrastive pair (a red and a purple butterfly) and a distractor (a red fox); or it contained a second contrastive pair (a red fox and a grey fox). The utterances further varied in focus, which was realised either through a cleft construction (literally: ‘red put butterfly’ vs. ‘red butterfly put’) and/or through pitch accent placement (early: on ‘red’ or late: on ‘butterfly’). Russian adults rapidly used the contrastive pitch on the adjective for anticipating the referent of the ensuing noun ‘butterfly’, and they were even faster when the visual context contained only one contrastive object pair (the red butterfly and the purple butterfly). The children, by contrast, waited for the noun ‘butterfly’ before inspecting the target, even when both the early pitch accent on the sentence-initial adjective and the visual context (one contrastive pair of objects) should have sped up their visual attention to the butterfly.

Another example of limited visual context effects in young children comes from an experiment in which Trueswell et al. (1999) examined whether young children (5-year-olds) rely on the so-called ‘referential principle’ for disambiguating a local structural ambiguity. In the sentence “Put the apple on the towel…” the prepositional phrase “on the towel” can either modify “the apple”, indicating its location, or the verb phrase, specifying the destination of the action. Adults prefer the destination interpretation (and inspect an empty napkin to which the apple can be moved) but rapidly switch to the location interpretation (inspecting the apple located on the towel) when the referential context supports this (e.g., when one of two apples is on a towel, Spivey et al. 2002; Tanenhaus et al. 1995).

Five-year-olds (4.8–5.1 years) by comparison did not apply this referential principle online. When listening to “Put the frog on the napkin in the box”, they frequently inspected a potential destination of the action (an empty napkin) instead of the frog on the napkin even when the context supported the location interpretation (one of two frogs was on a napkin). They did not even revise their initial analysis when executing the instruction, and put the frog onto the napkin (instead of into the box) on 60% of the trials. The authors concluded that children were not able to revise their initial syntactic analysis and suggested their more limited processing capacities relative to those of the adults may lead them to discard

---
3When more than one syntactic analysis is possible, the referentially supported analysis is preferred.
infrequent alternative structures and interpretations early on in favour of the dominant (destination) interpretation (Trueswell et al. 1999). A capacity account seems plausible since children can employ the referential principle provided they are given more time. Three-to-five-year-olds exhibited adult-like performance in a post-sentence act-out task (i.e., they moved the frog that was on the red napkin into the box) when they had listened to a similar sentence with their eyes closed (Meroni and Crain 2003).

Alternatively, children must engage in pragmatic processing to apply the referential principle. They must know which objects are present, notice the initial referential ambiguity (there are 2 frogs), and infer that “on the napkin” could disambiguate reference. At age 5, however, they may not yet be able to compute pragmatic inferences during realtime language comprehension. This would be in line with the observed delay in 5–9-year-olds computation of pragmatic implicature (Huang and Snedeker 2009a).

Yet another possibility is that the experimental stimuli discourage children’s reliance on the visual context (Zhang and Knoeferle 2012). When children heard “… on the napkin…” they could either inspect the empty napkin (a potential referent) or the frog located on the napkin. It is not implausible to assume that children pursue primarily a referential strategy (mapping “napkin” onto its referent), leading them to inspect the empty napkin (vs. another object on a napkin). If so, then children would be garden-pathed even more since a referential strategy would guide their attention to the incorrect destination (the empty napkin). If this were the reason for children’s failure to rapidly rely on the visual context, the results would no longer reflect an inability to rapidly rely on the visual context but rather garden-pathing precisely because children rely on the visual context (relating nouns to objects).

Indeed, when Zhang and Knoeferle (2012) examined visual context effects in a situation when a sentential verb referred to a depicted action event, the events rapidly affected 5-year-olds’ thematic role assignment and syntactic structuring. The events depicted, for instance, a bear pushing a bull and a worm painting the bear (Fig. 4.1). As they inspected this scene, adults and 5-year-olds listened to either an unambiguous subject-verb-object (SVO, “Der Bär schubst sogleich den Stier”, ‘The bear (subj) pushes soon the bull (obj)’) or an object-verb-subject (OVS) sentence (“Den Bär malt sogleich der Wurm”, ‘The bear (obj) paints soon the worm (subj)’). Adults anticipated the event agent (the bull) upon hearing ‘pushes’ (Zhang and Knoeferle 2012). Five-year olds exhibited a strikingly similar eye-movement pattern. They also anticipated the agent of the painting event more when the verb was ‘paints’ than ‘pushes’, and they inspected the patient of the pushing event more when the verb was ‘pushes’ rather than ‘paints’.

In children, this gaze pattern emerged later (during the post-verbal adverb) than for the adults, but still before the second noun phrase disambiguated the correct role filler. The similarities in the eye-movement pattern for the children and the adults suggest that 5-year-olds can rely on the non-linguistic visual context

---

4Object-initial word order is grammatical in German but non-canonical, and elicits processing difficulty (e.g., the case-marking on the determiner of the noun phrases elicits longer reading times).
(depicted events) for thematic role assignment and syntactic structuring. Note, however, that the sentences were structurally unambiguous. It remains to be seen whether depicted events also help children to rapidly disambiguate local structural ambiguity (Zhang and Knoeferle in progress).

Five-year-olds resembled adults also in their preference to rely on recently inspected events. Adults preferentially ground their spoken language comprehension in recent compared with future events. They inspected a clipart scene depicting a waiter polishing candelabra and moving away from them (Knoeferle and Crocker 2007, Experiment 3). Other objects (crystal glasses) also afforded polishing. An ensuing spoken German sentence referred either to the recent event (literal translation: ‘The waiter polished recently the candelabra’) or to a potential future event (literal translation: ‘The waiter polishes soon the crystal glasses’). The initial noun and the verb (up to the last two letters) did not reveal the tense of the event (past vs. future). Adults made more inspections to the target of the recently depicted than the future event (e.g., polishing the crystal glasses), a preference which generalised to real-world stimuli (Knoeferle et al. 2011a), and to experiments in which the future events were much more frequent than the recent events (Abashidze et al. 2013).

When two further participant groups (5-year-olds and adults) were each shown similar stimuli, analyses replicated the results from Knoeferle and Crocker (2007, Experiment 3). Participants saw a red barn, and a horse depicted as galloping to a blue barn. They listened to German sentences such as ‘The horse galloped to the blue barn’ (Zhang et al. 2012). The adults invariably inspected the blue rather than the red barn during the verb. The children exhibited a qualitatively similar but subtly delayed pattern, and inspected the blue barn post-verbally.

4.5 Summary

Not surprisingly infants and adults differ in how rapidly they relate language to the “world”. Infants often lag somewhat behind adults in their language-mediated eye-movement behaviour. This lag may reflect younger children’s limited processing capacity under the assumption that processing capacity affects the speed of language processing (e.g., Just and Carpenter 1992). Clear differences emerged regarding pragmatic and inferential processes, whereby adults but not children computed pragmatic inferences online (but see Papafragou and Musolino 2003).
With regard to language-world mapping, however, even young infants’ language-mediated visual attention behaviour resembled that of adults, and the literature on language processing documents their emerging ability to rapidly relate an unfolding utterance to objects in the visual context. While 5-year-olds may not always rely on the visual referential context, visual cues rapidly affected their thematic role assignment and syntactic structuring when mediated directly by language (e.g., a verb referring to an action). Five-year-olds also resembled adults in their preferred inspection of a recently depicted (over another future) action target but their eye-movement response was subtly slower than for adults.

4.6 Conclusions

This chapter has argued for an approach that strives to include a model of the comprehender into situated language processing accounts. To illustrate this approach I have discussed age-related variation in visually situated language processing. I have argued that visually situated language comprehension in young children resembles this process in adults. To corroborate this point, I have reviewed empirical research on (a) learning about language, objects and actions and on (b) language processing.

Insights from language learning can inform predictions for language comprehension. The literature on visually situated language comprehension suggests that contrast between objects can not enable rapid syntactic disambiguation in young children. This could be seen at odds with the important role of noun-object correspondences for language learning. One possibility is that certain processes such as syntactic disambiguation are exempt from visual context effects in child comprehenders. However, if we assume a similar role of visual context for language learning and comprehension, then object-based information should rapidly affect language comprehension. Assume, for instance, that children see a visual context in which one dog sits on a cushion while another dog sits on a rock. The children listen to “Den Hund auf dem Felsen füttert gleich der Mann.” (literal translation: ‘The dog on the rock (obj) feeds soon the man (subj)’). Object-first sentences are difficult to understand, but can be felicitous when they establish reference through contrast (‘on the rock/cushion’). If children can rapidly use that contrast for syntactic structuring, then they should anticipate the upcoming agent (the man) earlier in a context with two dogs than with one dog (and another animal). To the extent that these and related predictions are borne out, they would corroborate that a comparison of language learning and comprehension can help us to gain insight into age-related variation of visually situated comprehension, with the goal of establishing a model of the comprehender.

Acknowledgments This research was funded by the Cognitive Interaction Technology Excellence Center 277 (CITEC, German Research Council).
References


Zhang, L., & Knoeferle, P. (in progress). Can depicted events disambiguate local structural ambiguity in German SVO / OVS sentences?

