Theory of mind in language, minds, and machines: a multidisciplinary approach

Dijk, B.M.A. van

Chapter 6

Analysing Semantic Development with a Language Model

In this chapter we employ a Language Model (LM) to gain insight into how complex semantics of Dutch Perception Verb (PV) *zien* ('to see') emerge in children. Using a Dutch LM as representation of mature language use, we find that for ages 4-12y 1) the LM accurately predicts PV use in children's freely told narratives; 2) children's PV use is close to mature use; and 3) complex PV meanings with attentional and cognitive aspects can be found. Our approach illustrates how LMs can be meaningfully employed in studying language development, hence takes a constructive position in the debate on the relevance of LMs in this context.

This work was originally published as: Van Dijk, B.M.A., Van Duijn, M.J., Kloostra, L., Spruit, M.R., and Beekhuizen, B.F. (2024). Using a Language Model to Unravel Semantic Development in Children's Use of a Dutch Perception Verb. In Zock, M., Chersoni, E., Hsu, Y., and De Deyne, S., editors, *Proceedings of the 8th Workshop on Cognitive Aspects of the Lexicon*, pages 98-106. European Language Resources Association.

6.1 Introduction

Recent Language Models (LMs) based on Transformer architectures (Vaswani et al., 2017) reflect semantic knowledge present in a language community. BERT vectors (Devlin et al., 2019), for example, are able to distinguish different senses of the same word (Rogers et al., 2020; Vulić et al., 2020; Wiedemann et al., 2019). These LMs implement the distributional hypothesis that words with similar meanings tend to occur in similar contexts, and they represent both word type and word token meanings with real-valued vectors (Lenci and Sahlgren, 2023). The latter allows LMs to encode polysemy and different usages of words.

Despite this, LMs' relevance in the context of language development is disputed: their architecture and volume of training input have been argued to make them incomparable to children (e.g. Bunzeck and Zarrieß, 2023; Prystawski et al., 2022; Warstadt and Bowman, 2022). Yet, others argue that LMs can show which linguistic phenomena are *in principle* learnable from distributional information, bearing on learnability debates (Contreras Kallens et al., 2023; Piantadosi, 2023; Wilcox et al., 2023).

Here we leverage LMs' rich semantic information to gain insight in children's semantic and pragmatic development. Addressing the question whether children's pragmatic use of lexical items develops over time or, conversely, is adult-like from the start, we use a Dutch LM as representation of mature language use and study the Dutch Perception Verb (PV) *zien* ('to see'). We find that children's use of *see* is close to mature use across the 4-12y age range, and that for all ages the familiar mature usage patterns of the verb can be identified. As such, this chapter further illustrates the relevance of LMs in studying language development, by reflecting on LMs as representations of mature language use and setting up appropriate tasks and metrics.

6.2 Background

Little empirical work employs modern LMs in language development, the exception being work comparing word acquisition in children and LMs (Chang and Bergen, 2022; Laverghetta Jr and Licato, 2021). This is understandable given the debate on the validity of LMs in the child context: LMs and children differ in key respects including word exposure (Warstadt and Bowman, 2022) and learning mechanisms (Bunzeck and Zarrieß, 2023).

Still, LMs are arguably useful representations of *mature language use* by being

trained on corpora of adult language, and are therefore of value in modelling language understanding. LMs can be viewed as an incremental methodological step compared to earlier corpus studies comparing children's verb use to mature use, that relied on manual annotation or feature engineering to identify different senses of mature verb use (e.g. Adricula and Narasimhan, 2009; Parisien and Stevenson, 2009), but different senses, as we will show, can also be conveniently retrieved from LMs. These and other considerations have led to increasing acknowledgement of LMs' relevance for analysing language development (Contreras Kallens et al., 2023; Lappin, 2023), and efforts to make LMs more comparable to the child context (Warstadt et al., 2023).

Here we address the relevance of LMs in the developmental context by analysing children's lexical semantic development with LMs. We target children's use of Dutch PV *zien* ('to see') as a case study, which has been frequently analysed in language development (e.g. Davis, 2020; Davis and Landau, 2021). Studies of perception verbs across languages have shown that visual perception verbs have extended meanings beyond their denotational meaning 'entity X visually perceives object or event Y', that involve additional aspects of e.g. *attention* ("Let's see if I can find the keys") and *cognition* ('I see what you mean') (San Roque et al., 2018; San Roque and Schieffelin, 2019). Such meaning extensions are salient for children with a limited lexicon, where meaning extension of known words allows children to express new meanings efficiently (Nerlich and Clarke, 1999). In addition, since visual perception is argued to have strong metaphorical mappings to knowledge and understanding (e.g. Johnson, 1999), *see* can be a window onto how children learn to represent (socio-)cognitive content with language (Sweetser, 1990).

This work addresses the question of when meaning extension occurs. Some argue that literal understandings of PVs emerge first in young children (e.g. Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009), while others argue pragmatic meanings are likely present early due to the social situatedness of language learning (e.g. Enfield, 2023; San Roque and Schieffelin, 2019). In the latter case, the discursive relation between the visual perception event and the events surrounding it may be more salient for a language learner than the encoding of visual perception per se. For example, a young child's utterance *see ball* may be followed by the caregiver showing the ball, or focusing its attention on the ball — further attentional aspects that are likely relevant components of the message for the child beyond the denotational content of visual perception having taken place. While focusing on a single verb may seem limited, we believe as a case study, visual perception verbs are well-chosen as a starting point for generalising the proposed approach,

since their acquisitional pathway and pragmatic usages (as described above) are well understood.

We focus on children's use of *see* in ChiSCor, a corpus of freely told stories by Dutch children (4-12y) in classroom settings (for details see Chapter 3), since complex PV meanings can be especially relevant in the narrative domain. For example, that character X *sees* entity Y may not only imply that X literally perceives Y, but also that X *evaluates* Y or *discovers* Y. Such information, which may be crucial for the 'tellability' of the story (Labov and Waletzky, 1967), can be efficiently transmitted through PVs. Narratives are 'natural' sandboxes for children to challenge their language competence in various ways (Frizelle et al., 2018), including the development of lexical pragmatics.

6.3 Methods

Language data

We extracted all 308 occurrences of *see* from 619 stories of 442 children (4-12y) in ChiSCor. We manually inspected these occurrences and removed unintelligible usages (mainly transcription errors) as well as stories exceeding a context window larger than 512 tokens, resulting in 210 occurrences. We assigned occurrences to a Young (4-6), Middle (6-9) or Old (9-12) age group, following the age binning in Dutch primary education, and included only PV occurrences from one story per child, resulting in 30 Young, 82 Middle and 42 Old PV occurrences. To balance the sample across age groups, we randomly sampled 30 occurrences from the Middle and Old age group.

A known problem with LMs is that data contamination can lead them to solve tasks by memorisation (Deng et al., 2024). ChiSCor is likely not in the train data of recent LMs, as the corpus is recent and 'hidden' behind view-only links in research papers. Further, ChiSCor's free storytelling is unlike other available Dutch corpora that involve language elicitation and as such constitutes language that tests LMs' generalisation capabilities.

LMs as benchmark models

Using LMs as representation of mature language use requires evidence that the LM models the linguistic phenomenon and domain at issue reliably. We draw on findings that word representations in BERT encode rich semantic information about word polysemy (Garí Soler and Apidianaki, 2021; Wiedemann et al., 2019), although not per-

fectly. Also, Dutch LMs are for a large part trained on narrative texts (e.g. De Vries et al., 2019; Delobelle et al., 2020), and LMs in general have been shown to model coherence in written narratives well (Laban et al., 2021). In sum, earlier work supports the idea that LMs encode mature PV use in narratives.

Choice of LMs

For reasons of computational efficiency, validity with respect to the child context, and reproducibility, we chose `RobBERT-2023-dutch-large`, a Dutch BERT-like LM (Delobelle et al., 2020). RobBERT has 455M parameters trained on 19.5B tokens and is more in line with the 100M token training input a 10-year-old has seen (Warstadt and Bowman, 2022), compared to often employed larger LMs like GPT-3 (175B parameters, 500B tokens (Brown et al., 2020)).[1] RobBERT is accessible through the `HuggingFace Transformers` ecosystem (Wolf et al., 2019).

Recent work on LM relevancy to human language acquisition in the BabyLM challenge (Warstadt et al., 2023), highlighted smaller LMs with optimised architectures and train objectives, and curated train data for training developmentally plausible models (Samuel et al., 2023). However, such Dutch LMs are not yet available and training models from scratch is generally not feasible for researchers studying language acquisition. RobBERT was a fitting resource as it is optimised compared to BERT and has a simpler training objective (masked language modelling only) (Liu et al., 2019). These aspects go some way towards the findings of the BabyLM challenge (Samuel et al., 2023; Warstadt et al., 2023).

Task design and metrics

To use LMs as representations of mature language use, zero-shot evaluation settings as described by Laban et al. (2021) are preferred. This means using LMs of-the-shelf without further pre-training on the target domain or fine-tuning to stay close to the mature language use encoded in the LM, similar to how factual knowledge can be retrieved from LMs without fine-tuning (Petroni et al., 2019). We use various possibilities available through LMs to assess whether and how children's use of *see* differs from mature use.

Our first task consists of predicting *see* in children's narratives. We present RobBERT with stories containing a masked instance of *see*, as in the (translated) excerpt

[1] In the context of this chapter scale differences between BERT and GPT-3 are most salient, but we acknowledging that GPT-3 as unidirectional decoder-only model is also qualitatively different from BERT-like models like RobBERT.

in (1):

(1) [...] one time robot was travelling. and all of a sudden he <**mask**> a wolf. and he ran away quickly. [...] (Story ID 052301)

In our experiment we provided full stories as context to RobBERT, which varied in number of words ($\bar{x} = 187, \sigma = 108$). If children's usage differs from adults, the LM might have difficulty predicting the PV correctly.

As a second measure, we compute the negative log-likelihood NLL or surprisal for a prediction for a masked token w_m with

$$NLL(w_m) = -\log p(w_m | w_{1\ldots m-1}, w_{m+1\ldots n}) \tag{6.1}$$

with the `fill-mask` pipeline from `HuggingFace Transformers`. This measure provides further context to the predictive accuracy measure presented above: lower NLL implies that the predicted token is less surprising i.e. closer to mature use as encoded in the LM, and more generally indicates how well a given context supports a specific token on the masked position (PV or other).

Lastly, we use the tokens in RobBERT's top-5 predictions for masked instances of *see* as 'near neighbours' that can reveal the additional discursive meanings that the usage of PVs supports. Our data and notebooks are available at `https://osf.io/8eyvf/`.

6.4 Results

Predictive accuracy

First, we assessed RobBERT's overall performance in predicting *see* at masked positions in all 90 PV occurrences. Accuracy is overall high (.83, Table 6.1), and although lower for Young (.70) we found no significant difference in accuracy between ages with an ANOVA ($F_{2,87} = 2.974, p = .056$).

This shows that RobBERT models children's PV use in the narrative domain well. The 15 errors were mainly in Young and showed confusion of *seeing* with 'finding', 'having', 'looking' and 'getting', meaning that contexts underconstrained the use of *see*. Although these other verbs can be valid tokens on masked positions (e.g. 'found' in (1)), here our aim was to see if RobBERT adequately models that *see* can subsume such other possible meanings in narratives.

Metric	Young	Middle	Old	Overall
Accuracy	.70 (30)	.90 (30)	.90 (30)	.83 (90)
Surprisal	.40 (21)	.23 (27)	.32 (27)	.31 (75)
Top-5	1.00 (30)	1.00 (30)	.97 (30)	.99 (90)

Table 6.1: Metrics for RobBERT. Accuracy: percentage that *see* was predicted. Surprisal: NLL computed for predictions of *see*. Top-5: proportion that *see* was in top-5 predictions. Number of PV occurrences (i.e. observations) in parentheses.

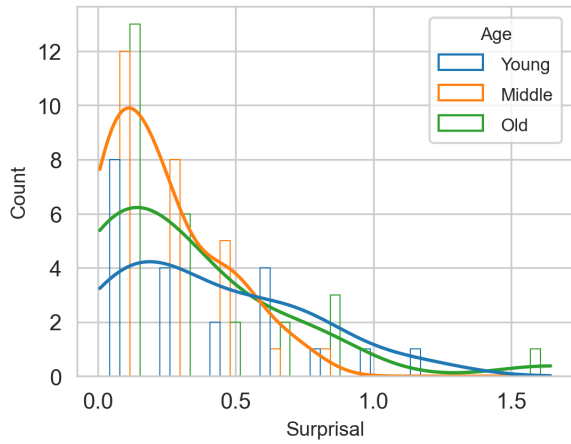

Figure 6.1: Surprisal distributions.

Surprisal

Second, we analysed potential age effects in mean surprisal for 75 correct predictions of *see*. For example, RobBERT may be less surprised by PV use for Old compared to Young or Middle, indicating that PV use of Old children is closer to mature use than for Young. Interestingly, surprisal distributions tend to 0 for all ages (Figure 6.1), suggesting that use of *see* is overall close to mature use irrespective of age. And although mean surprisal between Young, Middle, and Old differs (Table 6.1), pairwise comparisons with Tukey's HSD (Tukey, 1949) revealed no significant age effects. This indicates that PV use by children of all ages is about equally close to mature use as approximated by RobBERT.

Top-5 alternative predictions

For virtually all age groups, *see* is in the top-5 predictions (Table 6.1), which supports the idea that by examining top-5s we get insight into extended meanings of *see*. For

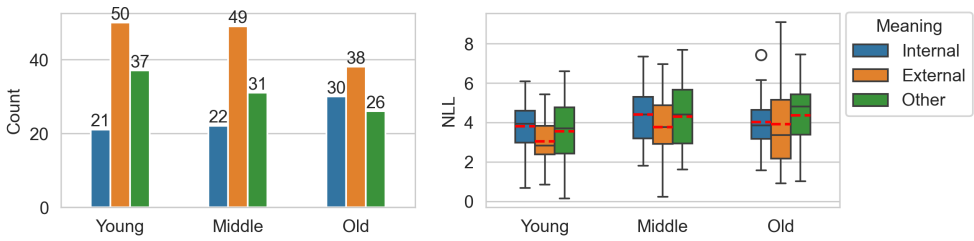

Figure 6.2: Frequencies (left) and surprisal dist. (right) of internal, external, and other meanings of 304 top-5 lemmas. Bars (left) stack to 100%; dashed red lines (right) indicate means.

90 PV occurrences and their top-5s (450 tokens) we lemmatised tokens and removed *see* and lemmas that were not verbs (e.g. 'many', 'and', 'at'), resulting in 304 lemmas. We then took the set and classified 65 lemmas as having roughly `external`, `internal`, or `other` meaning. `External` implies a meaning pertaining to plain action (e.g. 'to go', 'to come', 'to carry', 'to throw'); `internal` a meaning pertaining to an attentional (e.g. 'to notice', 'to meet') or cognitive state (e.g. 'to think', 'to know'). `Other` pertains to auxiliary verbs and PVs not the focus of the current study (e.g. 'to have', 'to hear').

The idea is that top-5 lemmas indicate what possible meanings PV contexts support, even if these lemmas are not necessarily intuitive substitutions. For example, substituting 'threw' for <**mask**> in (1) renders the excerpt less intuitive. Yet, this immediate context as a sequence of *external* actions better supports understanding *seeing* also as a causal part of a sequence of external actions, than as *seeing* as part of narrative components reflecting a character's attentional or cognitive *internal* states (cf. examples in Table 6.2).

We assessed frequencies of `external`, `internal` and `other` meanings, and their mean surprisal over age groups to identify potential age differences in occurrence and closeness to mature use. Regarding frequency, although `external` and `other` meanings decrease over age while `internal` meanings increase over age (Figure 6.1, left), we found no significant age effects with a χ^2 test of independence $\chi^2(4, N = 304) = 5.044, p = .283$, suggesting that all the different meanings are about equally frequent in Young, Middle and Old groups. Regarding surprisal (Figure 6.2, right), distributions for `external`, `internal` and `other` meanings are relatively similar both within and between age groups. Pairwise comparisons with Tukey's HSD found only a significant difference at the $p < .05$ level between mean surprisal for `external` meanings for Young and Old.

Age	Ex.	PV context
Young	**(2)**	*.. and when he returned. then he saw/knew that the princess was gone. and they lived happily ever after.* (102901)
	(3)	*.. and then they were lost again. and then they saw/searched the castle. and then they went in the castle.* (122901)
	(4)	*.. but then the teacher came and then she was already too late. the teacher had seen/caught them. and then you get a punishment from the teacher.* (033401)
Middle	**(5)**	*.. but then they lost each other all of a sudden. and then Wergje saw/met another rabbit. and it asked how are you called.* (072301)
	(6&7)	*.. because when he was home. then he saw/noticed/discovered that he had the other scales. but then he went to fly on it and he wanted to find his own dragon again.* (022301)
Old	**(8)**	*.. once arrived at the cave Puta completely forgot that you were not allowed to touch the big diamond. Puta saw/checked out the diamond and found it so beautiful. and he touched it accidentally.* (034801)
	(9)	*.. so then the fat little king went on his fat broom to the cry for help. and what did he see/think. the cry came from a little fat guinea pig that looked very much like the king.* (023801)
	(10)	*.. and he ever wanted one time to try it with his eyes closed. to see/test can I grab that donut well with my eyes closed.* (034501)

Table 6.2: Translated PV contexts with top-5 internal lemmas (underlined) with lowest surprisal. Story IDs given in parentheses. All excerpts were translated from the original stories.

We illustrate complex meanings of *see* present in all age groups, by providing the three `internal` meanings that were closest to mature use (i.e. with lowest surprisal) and their PV contexts in Table 6.2. We make three observations. First, `internal` meanings with attentional and cognitive aspects can be but are not exclusively cued by surface linguistic frames such as complementation that RobBERT simply picks up, as example (4) and (9) show. In (4) 'caught' implies that the teacher knows what the 'she' character is up to; in (9) 'think' renders the realisation where the cry of help is coming from a representation in the mind of the king. Second, `internal` meanings are varied: from more purely attentional where characters simply become aware of something or find something out as in (6&7), to more social (5), and evaluative attentional aspects (8). Third, although `internal` meanings with cognitive aspects have the most abstract lemmas ('think', 'know') that are argued to be harder to master (Barak et al., 2012), cognitive meanings were found in both Young (2), (4) and Old (9) children.

6.5 Discussion

Our results show that complex meanings of the Dutch perception verb *zien* ('to see') are about equally frequent in all age groups and that children's use of the PV is overall not significantly different from mature use. This contrasts with earlier work that has argued that children initially acquire more literal meanings of PVs (Adricula and Narasimhan, 2009; Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009) (Section 6.2), although we note that children in our sample are older (four years and older) than children in earlier studies (typically between two and four years).

Our result aligns with the idea that it is the social context that cues various complex senses of *see* in children (e.g. Enfield, 2023; San Roque and Schieffelin, 2019), and with the idea that (young) children may employ PVs like *see* as linguistic devices for learning to represent cognitive and attentional states (Johnson, 1999; Sweetser, 1990). We argue that our finding can be explained by the social context provided by live storytelling. PVs like *see* are linguistic devices for efficiently communicating about characters' attentional and cognitive states that are key to understanding the story, as PVs can compress redundant information that would make the story tedious. An earlier chapter has shown that in children's live storytelling, contexts of PVs like *hear* and *see* are coherent and clear, as evidenced by the rich PV vectors that can be trained from limited amounts of narrative data (Chapter 3).

Narrative language data may explain the contrast between our and earlier findings, as storytelling has been argued to solicit 'maximal behaviour' in that it challenges children's linguistic competence (Frizelle et al., 2018; Southwood and Russell, 2004), more than the speech produced by children in child-caregiver interactions would do, which typically take place in mundane contexts. Some earlier work contrasting with our results relied on language data from such child-caregiver interactions (e.g. Adricula and Narasimhan, 2009; Davis and Landau, 2021). The latter work also employed smaller sample sizes with less unique children and more PV use per child compared to the current study, which may compress the variation in complex semantics we find in our analysis.

Interestingly, RobBERT accurately predicted *see* in narratives of children of all ages; we argue that this is not a mere frequency effect (i.e. *see* being more frequent in train data than alternatives), given that top-5 predictions often reveal RobBERT's correct mapping of the nuanced senses of PVs. Also, RobBERT's aptitude in handling PV use in narratives is interesting insofar children's stories are not obvious regarding

wording, characters and themes. One issue pointed out by a reviewer is whether LMs with Transformer architectures are the best fit for representing linguistic knowledge of a mature Dutch language user, or whether other models should be used, e.g. from the BabyLM challenge (Warstadt et al., 2023). The best-performing LMs in this challenge employed Transformer architectures that are essentially optimised versions of vanilla BERT models regarding training objective, architecture and dataset (Samuel et al., 2023). With our choice for RobBERT we aimed to make the comparison to the human case as valid as possible with an existing resource (see Section 6.3).

In any case, from the BabyLM challenge we learn that the Transformer architecture is also in more modest training setups a powerful encoder of linguistic information. Our claim is not that Transformers are therefore good (cognitive) models of human language users, which is still debated (see e.g. Paape, 2023, and Chapter 8). Rather, when it comes to specific linguistic aspects such as mature semantic and pragmatic knowledge, LMs as sophisticated distributional learners represent this information in a convenient fashion. For using such computational models as representations of mature language use, the primary question is if their *behaviour* for a specific linguistic phenomenon is sufficiently complex, which for many modern BERT-like models seems the case. But representations of mature use could also be created in other ways, e.g. by clustering different verb senses with features based on verb argument structure in a large corpus of mature language use. Thus, LMs are more of an analytical tool here than direct models of humans. That said, it is still worthwhile and necessary to make LMs more similar to the human context.

6.6 Conclusion

This chapter provided a case study on Dutch children's (4-12y) use of *zien* ('to see') and the emergence of complex semantics in the use of this perception verb. We showed that 1) a recent Dutch LM can predict use of *see* in narratives produced by children of different ages reliably; 2) children's use of *see* is close to mature use for all ages; and 3) complex meanings of *see* with attentional and cognitive aspects can be found across all ages. Our results align with work that argues that meaning extension occurs early in children and with the idea that via perception verbs, children may learn to represent socio-cognitive content.

We also showed how LMs can be meaningfully leveraged in developmental contexts. We hope to provide future researchers with useful reflection on how to proceed when using LMs as representations of mature language use, choosing models, and

setting up tasks and metrics.

6.7 Limitations

A limitation of this study is that we provided the whole story as context for predicting a masked occurrence of *zien* ('to see'), but for space limitations we could only discuss complex meanings with smaller story excerpts as in Table 6.2. This may suggest that complex PV meanings can be determined from small pieces of narrative after all. Yet, when doing the same task with smaller PV contexts as in Table 6.2, i.e. a sentence before and after the sentence featuring an occurrence of *see*, RobBERT's overall accuracy drops from .83 to .57 and overall surprisal increases from .31 to .59 (see Table 6.1), which suggests that RobBERT needs to take the whole story into account to model PV use adequately. This means that there is more relevant information in the context beyond what we show in the immediate PV context that renders RobBERT's predictions of masked tokens accurate and supports additional meanings of *see*.

Another limitation is that we had to translate story excerpts into English, as also providing Dutch excerpts required too much space. Some awkwardness in translations could not be avoided. For example, Dutch has a verb 'betrappen' that always has a cognitive meaning similar to 'catching somebody red-handed', whereas 'catching' in English can also have a more obvious action-related meaning. 'Betrappen' was a token prediction in RobBERT's top-5 with low surprisal that we had to translate as 'caught' in example (4) in Table 6.2.

6.8 Ethics Statement

In this study we used the ChiSCor story corpus and we refer to Chapter 3 for further details regarding ethical considerations and approval that was obtained for collecting language data from children. Regarding computational efficiency, we chose a relatively small, open and free to use Large Language Model that can also be employed with limited computational resources.