Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5198–5205
Marseille, 20-25 June 2022

ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus

Francesco Cabiddu[1], **Lewis Bott**[1], **Gary Jones**[2], **Chiara Gambi**[1]
[1]Cardiff University, [2]Nottingham Trent University
{cabidduf, bottla, gambic}@cardiff.ac.uk, gary.jones@ntu.ac.uk

Abstract

Language acquisition research has benefitted from the use of annotated corpora of child-directed speech to examine key questions about how children learn and process language in real-world contexts. However, a lack of sense-annotated corpora has limited investigations of child word sense disambiguation in naturalistic contexts. In this work, we sense-tagged 53 corpora of American and English speech directed to 958 target children up to 59 months of age, comprising a large-scale sample of 15,581 utterances for 12 ambiguous words. Importantly, we carefully selected target senses that we know - from previous investigations - young children understand. As such work was part of a project focused on investigating the role of verbs in child word sense disambiguation, we additionally coded for verb instances which took a target ambiguous word as verb object. We present experimental work where we leveraged our sense-tagged corpus ChiSense-12 to examine the role of verb-event structure in child word sense disambiguation, and we outline our plan to use Transformer-based computational architectures to test hypotheses on the role of different learning mechanisms underlying children word sense disambiguation performance.

Keywords: Word Sense Disambiguation, Verb-Event Structure, Child-Directed Speech

1. Introduction

Although theories of word learning predict that young children do not map word forms to multiple meanings (Markman, 1989; Trueswell et al., 2013), recent evidence suggests that the speech children hear early in development is rich in word sense ambiguity, and also that children's early vocabularies are populated by ambiguous words (Meylan et al., 2021). This raises a number of questions: Which environmental factors and which learning mechanisms might help children deal with sense diversity in real-life contexts? Developing annotated corpora is crucial to answering such questions: developmental researchers can test the effect of naturalistic variables by constructing experimental stimuli based on corpus statistics, and they can implement and test the plausibility of hypothesized learning mechanisms by using computational architectures trained on naturalistic conversations (Monaghan and Rowland, 2017).

In this work, we focus on the role that verbs might play in word sense disambiguation during sentence processing. Verbs play a key role in different aspects of language acquisition (unambiguous word processing, Mani, Daum, and Huettig, 2016; syntactic ambiguity resolution, Kidd, and Bavin, 2005), but their role has received less attention in child word sense disambiguation. To help fill this gap, we share a large corpus of child-directed speech, ChiSence-12, where we tagged all verb-sense occurrences of 12 ambiguous words used in previous developmental studies. The corpus can be freely downloaded at https://gitlab.com/francescocabiddu/chisense-12. We present preliminary analyses of experimental work based on ChiSense-12, and outline our plans to train Transformer-based computational architectures on this corpus to test hypotheses about the learning mechanisms involved in learning ambiguous words.

2. Related work

2.1 The role of verb-event structure in child word sense disambiguation

A key question in language acquisition research concerns the type of information sources children rely on in sentence parsing, which also constrains the type of learning mechanisms involved in such tasks (e.g., Ambridge, Pine, and Lieven, 2014). This question is important for theories that assume young children rely only on low-level information when parsing speech (e.g., statistical regularities; Snedeker and Yuan, 2008), and for those that allow additional integration of high-level information when this is judged as sufficiently reliable (e.g., syntactic or semantic structures; Trueswell and Gleitman, 2007).

In the context of word sense disambiguation, reliance on statistical regularities hinders correct parsing under conditions where low-level and high-level cues are put in competition (e.g., Khanna and Boland, 2010; Rabagliati, Pylkkänen, and Marcus, 2013). For example, in Rabagliati et al. (2013), 4-year-olds presented with the spoken story *Elmo watched a funny movie about a castle, and a princess, and a silly dragon. That was a funny night* struggle to use the sentence global coherence (e.g., people usually watch movies at *night*) and tend to interpret the final noun as the homophone *knight*, given its statistical association (in naturalistic speech) with *castle*, *princess*, and *dragon*.

Although such evidence seems to support the idea that children are blind to high-level factors, other cues to sentence parsing might gain a status of reliability early in development. For example, there is indication that at least from the second year of age children exploit verb-event structure when using sentence context to process unambiguous word meanings: for example, even if they have never heard the expression *pushing a flowerpot* in conversation, they know it is more plausible than *pushing a road* (Andreu, Sanz-Torrent, and Trueswell, 2013; Mani et al., 2016). In other words, verb-event structure refers to semantic restrictions that verbs impose on their arguments. Evidence also suggests that children might be able to rely on verbs to resolve lexical ambiguities, with 4- to 7-year-olds being able to understand that, for example, *She met the star* refers to *star[famous person]* and not *star[astronomical object]* (Hahn, Snedeker, and Rabagliati, 2015; Rabagliati et al., 2013). However, it is still unclear whether facilitation from verbs comes from low-level associations (e.g., in the language *meet* co-occurs with *star[famous person]* but not with *star[astronomical*

object]) or from verb-event structure (e.g., one more plausibly meets animate objects).

In order to disentangle the effect of lexical associations and verb-event structure, we leveraged statistics extracted from a naturalistic corpus of child-directed speech to construct carefully designed evaluation measures. In the following paragraphs, we describe the procedure for annotating the corpus and present some preliminary experimental work. Then, in section 6 we outline our plans to use Transformer-based models to examine the developmental plausibility of domain-general learning mechanisms that might underlie child word sense disambiguation performance.

2.2 Sense-annotated corpora

Numerous sense-annotated corpora based on adult language exist (for an overview, see Pasini and Camacho-Collados, 2020). In these, sense annotation is usually based on Wikipedia pages or the sense inventory WordNet (Fellbaum, 1998). A supervised approach to Natural Language Understanding - which makes use of sense-annotated training corpora - has proven to be useful in capturing adults' word sense disambiguation. For example, Loureiro et al. (2021) have recently shown that Transformer-based Language Models more closely approximate adult sense inter-annotator agreement when trained on sense-annotated instances compared to uniquely exploiting glosses from sense inventories (or a combination of the two). Although this evidence highlights the potential of using adult corpora for adult word sense disambiguation, the same is not necessarily true for studying child competence.

Compared to adult language, speech that young children hear is more repetitive, restricted to certain topics and concrete vocabulary (e.g., food, clothing, animals), with shorter sentences and simpler syntactic structure (Saxton, 2009). These characteristics may play a key role in early word processing and learning (e.g., Weisleder and Fernald, 2013), indicating that experimental or computational investigations aimed at capturing children's language understanding should be based on the specific input they receive. Furthermore, sets of ambiguous words tagged in adult corpora may consider senses that are not understood by children, or conversely, they may omit senses that are understood by children. This makes it important to select samples of word senses that young children understand.

In the first work addressing these challenges, Meylan et al. (2021) are currently tagging two large corpora of English child-directed speech (and corresponding child productions) from the CHILDES database (MacWhinney, 2000), which is a large collection of conversational transcripts. The child-directed corpora comprise speech directed to 18 children of age between 9 and 51 months. A total of 112,802 word tokens is being tagged using WordNet sense inventory as a reference. The sample of word types considered are based on a common measure of child vocabulary from parental report, the Communicative Development Index (CDI; Fenson et al., 2007), covering a total of 719 lemma+part-of-speech combinations in the corpus.

Although Meylan et al.'s dataset will significantly contribute to the naturalistic study of lexical ambiguity in early childhood, it is less useful for examining the contribution of specific aspects of sentence context such as verb-event structure. First, tagging specific syntactic patterns (e.g., verb-object) was not the focus of the project. Secondly, high-frequency words in the dataset are downsampled (i.e., a random sample of 50 tokens in each 3-month recording interval is tagged) to minimize annotation time. Although this seems a reasonable strategy when focusing on word sense distributions for each word type, it limits the researcher's ability to look at the distribution of verbs that co-occur with each specific sense (i.e., the verb distribution becomes especially downsampled for senses that appear infrequently in the corpus). For this reason, we used a large English corpus of child-directed speech where we manually tagged the full sample of tokens for both word sense and verbs that take a sense as an object. Given the large-scale nature of the project, to make the annotation task manageable we only coded a pre-selected sample of words. We describe the corpus, the word sample, and our annotation strategy below.

3. Corpus

We downloaded all American and British English corpora from the CHILDES database (version 2020.1) using the R package *childesr* (Braginsky et al., 2019), which provides a standardized procedure for downloading transcripts from different corpora. Out of 72 corpora downloaded, we considered 53 involving target children of up to 4 years of age (59 months), resulting in speech directed to 958 target children. We further filtered the dataset for utterances containing 12 ambiguous words (see Table 1). For each word, a frequent dominant sense and a less frequent subordinate sense were considered (e.g., Bat: *dominant* = animal, *subordinate* = object). 11/12 words were selected from a previous study where 4-year-olds showed understanding of both dominant and subordinate senses (Rabagliati et al., 2013). An additional word was selected with both senses having a relatively high frequency in child-directed speech (/ˈflaʊə/: flower/flour). Note that at least 90% of caregivers of children who took part in the pilot study presented in Section 5 reported their child understood both senses of this word.

In general, the sample of words was selected based on whether the distribution of verbs for each sense allowed to construct stimuli for the experimental conditions summarized in section 5.1.

4. Annotation

The dataset was tagged by the first author. We only considered utterances where a target word was used in its dominant or subordinate sense. Each utterance was tagged for the word sense used (dominant/subordinate). For utterances where the sense was used as object argument, we reported the verb stem preceding the sense (see Figure 1).

For utterances where the word sense was not immediately understandable, the surrounding conversational context was considered (i.e., surrounding utterances in the transcripts; see Figure 2). If the conversational context did not allow the annotator to understand the intended meaning, the utterance was discarded.

	A	B	C	D	E
1	ID	GLOSS	TARGET	SENSE	VERB
2	311504	who put the rubber band on there	band	object	put on
3	326153	are you in a marching band	band	music group	be in
4	326190	oh a clown's in the band	band	music group	be in
5	326293	remember Child when did we see a band	band	music group	see
6	326309	didn't we see a marching band	band	music group	see

Figure 1: Example of coded utterances. ID is the CHILDES database utterance number. This identifier can be used to retrieve specific corpus variables including speakers and target children's information. The remaining columns contain the target utterance (GLOSS), target ambiguous word (TARGET), specific word sense (SENSE) and verb stem used with that sense (VERB).

The final dataset included 15,581 utterances out of an initial raw sample of 21,342 (*word tokens* = 115,272; *word types* = 4,805). The dominant sense appeared on average 73% of the time (*SD* = 13%). Descriptive statistics for each ambiguous word are presented in Table 1.

Word (D/S)	*N* (D/S)	*Dominance*	*Length*
Band (Object/Music Group)	178/58	75%	4
Bat (Animal/Object)	247/130	66%	3
Bow (Knot/Weapon)	230/27	89%	2
Button (Electronic/Clothing)	568/285	67%	5
Chicken (Animal/Food)	1463/937	61%	5
Flower/Flour	3521/350	91%	4
Glasses (Eye/Drinking)	683/620	52%	6
Letter (Alphabet/Mail)	1446/946	60%	4
Line (Geometric/Row)	471/241	66%	3
Moose/Mousse	178/42	81%	3
Nail (Finger/Tool)	460/106	81%	3
Sun/Son	2029/365	85%	3
MEAN (*SD*)	-	73% (13%)	3.8 (1.1)

Table 1: For each target word, the table shows the raw number of utterances in which dominant (D) and subordinate (S) meanings appeared, percentage of utterances in which dominant sense appeared (Dominance), and length of word in phonemes (Length).

As there is an overlap of 8/12 words with Meylan et al.'s (2021) dataset, we plan to analyze inter-annotator agreement as soon as this large-scale dataset is released. To give an idea of the difficulty of the annotation task, we conducted a small inter-annotator agreement study, generating a random list of 45 sentences from the coded corpus (5 per target word, excluding target words that are not homographs, i.e., moose/mousse, flower/flour, sun/son). After a short training (using 5 training conversations), a second annotator read 45 test conversations between a child and one or more adults (see Figure 2). For each conversation, the second annotator was asked to indicate whether a target ambiguous word highlighted in red referred to its dominant meaning, subordinate meaning or to something else.

Mother: that's a different kind of wok
Mother: book
Target_Child: bat
Mother: should we turn the page
Mother: you know there's a zoo that we could go to this summer where they have bats
Mother: huh
Target_Child: no this
Mother: that's right
Target_Child: rrrr
Mother: a **bat**
Mother: do you see any cats on this page
Mother: vegetables
Mother: what are the things on this page
Mother: would you like to do that
Mother: rrrr oh
Target_Child: no this
Mother: it's called a bat
Mother: how about do you know what that one is

Figure 2: Example of test conversation in the small inter-annotator study. The target word in red is surrounded by its conversational context.

We found 100% agreement between first and second annotator (*Kappa* = 1, perfect agreement). The scripts for generating the random list of sentences and the small study results can be found in the project GitLab page.

5. Adults and Children study

5.1 Set-up and Hypotheses

We used our sense-tagged corpus to design an experimental set-up where we could test whether children as young as 4 years can use verb-event structure to resolve lexical ambiguities.

Participants were tested using an online forced-choice task (see Figure 4), in which they listened to spoken stories while looking at 4 pictures appearing on the screen (2 depicting the dominant and subordinate target senses, 2 distractors matched to targets for their frequency in the corpus and also representing good completions of the stories). Participants were then asked to select the image that goes well with last (ambiguous) word of each story. As shown in Figure 3, we constructed spoken stories consisting of a prior sentence context and a target context. The prior context was always lexically biased toward the subordinate meaning of a target ambiguous noun (*listen* and *music* in *Sophia listened to some music* co-occur often with *band[music group]* in the child-directed speech corpus).

Following the prior context, participants heard 1 of 3 target context conditions. In the control condition, the sentence verb frequently co-occurred with the subordinate target sense (e.g., caregivers often talk about *playing in a band[music group]*) and more plausibly accepts the subordinate sense as object argument (e.g., one more plausibly *plays in a band[music group]* than *band[object]*).

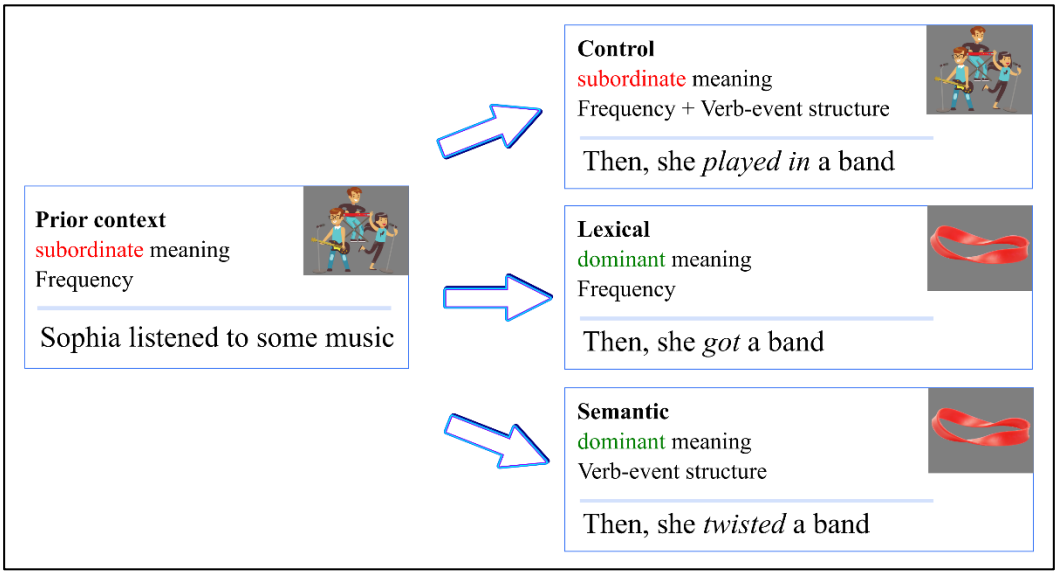

Figure 3: Example of trials in the 3 study conditions (control, lexical and semantic).

We consider this a control condition because it measures whether children can use a fully coherent sentence context (with congruent low-level and high-level cues) to resolve lexical ambiguities at all.

In two additional experimental conditions (lexical and semantic), we instead assessed whether children rely on low-level (lexical) or high-level (semantic) information carried by a local verb when the global context provided (conflicting) low-level information as cues to target sense. This is especially important for assessing reliance on verb-event structure, as implementing competition between cues excludes the possibility that children rely on verb-event structure only when this is the only cue available in sentence context (Rabagliati et al., 2013).

In the lexical condition, the sentence verb was lexically biased toward the dominant sense[1] (caregivers more often talk about *getting a band[object]* than *getting a band[music group]*), although both senses can be used as plausible arguments of the verb (i.e., neutral verb-event structure).

In the semantic condition, the sentence verb never appeared with any of the senses in the corpus. Also, the verb more plausibly accepts the dominant sense as object argument (one can more plausibly *twists a band[object]* than *band[music group]*).

As semantic continuations can be more difficult to process than controls - because one has to picture a more unusual scenario and make a higher number of inferences (e.g., *Olivia had some chips. Then, she rescued the chicken*) – we reduced the global plausibility of the control stories by avoiding any causal links between prior context and control sentences; instead, we used a temporal connective (*Then*) which is considered the lowest level of conceptual coherence save for completely unrelated sentences (see Connell & Keane, 2004; compare *Sophia listened to some music. Then, she played in a band* to *Sophia could not play the guitar, so she had to leave the band*).

As explained in section 2.1, given both the prominent role of lexical association in lexical ambiguity resolution (e.g., Rabagliati et al., 2013) and the key role of verbs in early word processing (e.g., Mani et al., 2016), we formulated the following hypotheses:

- Children will rely on verb-event structure (over lexical association from prior context). They will therefore select the dominant sense more in the semantic condition than control;

- Opposite lexical associations from prior context and verb-sense pairs in the lexical condition will show an additive effect, therefore children will select the dominant sense more in the lexical condition than control.

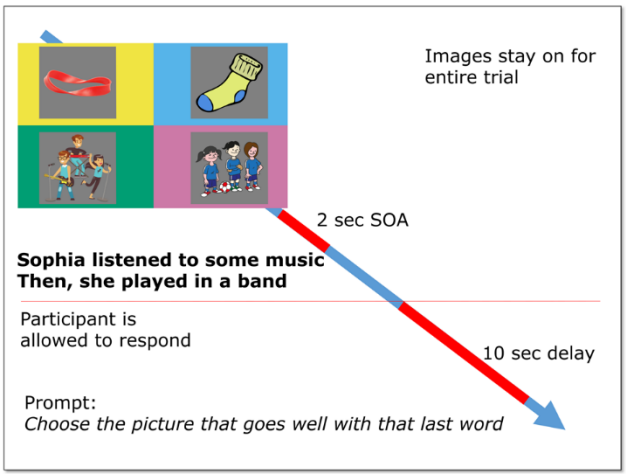

Figure 4: Example trial of control condition. Participants see a 2x2 grid displaying 2 alternative target senses (e.g., music band, elastic band) and 2 distractors (team, sock). After 2 seconds, a spoken story is played. Subsequently, participants are allowed to select the picture that goes well with the last word of the story.

[1] Verb-sense lexical association was computed as the verb-sense raw frequency weighted by the total number of times a sense appeared in the corpus as a verb object.

5.2 Participants

83 adult participants were recruited (*age*: *M* = 23 years, *SD* = 5 years; 62 women) using the Prolific platform. Adults were used to validate the experimental design and to make sure that the majority of participants named the study pictures using the labels we used for target-distractor frequency matching. We then conducted a power analysis via simulation to estimate the sample size for the children study (see pre-registration at https://osf.io/b87c6). Child data collection is currently ongoing, but we show preliminary descriptive statistics for the first 11 children tested so far.

5.3 Statistical analysis

We fitted a mixed-effect logistic model to adults' data, including image choice (dominant/subordinate) as the dependent variable, and condition (control, lexical, semantic) as the within-subject independent variable. We fitted a maximal random effect structure, which included participant and item type random intercepts, and random slopes of condition by participant or item type. We excluded estimated correlations between item random intercept and slopes because our simulations indicated insufficient power to detect the effect sizes of interest when such estimates are included.

5.4 Results

In Figure 5, we plot adults' proportion of image choice by item type. As expected, participants used the story context when this pointed toward the subordinate meaning (control). Moreover, when lexical association from prior context and verb-event structure from target context were put in competition (semantic), participants relied on verb-event structure and selected the dominant sense. The difference in dominant sense choice between semantic and control conditions was significant (*Odds Ratio* = 759.56 [231.61, 2491.00], $p < .001$).

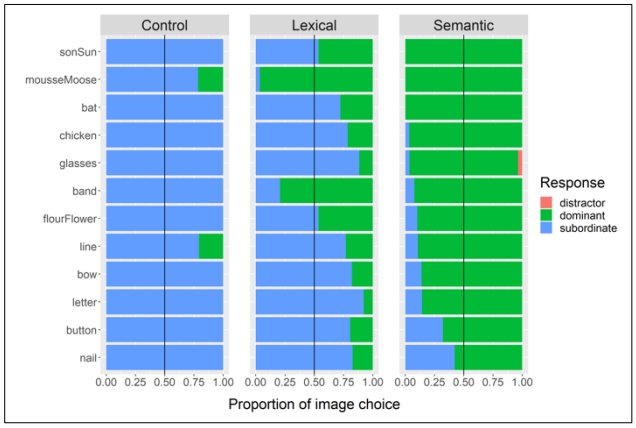

Figure 5: Proportion of image choice (distractor, dominant, and subordinate) by item type, in each study condition (control, lexical, and semantic). Vertical lines indicate .5 proportion of choice.

Adults also selected more dominant senses in the lexical condition compared to control (*Odds Ratio* = 25.29 [9.00, 71.05], $p < .001$). Given that the verb-event structure accepts both dominant and subordinate senses in the lexical condition (*Then, she **got** a band*), participants' choice toward the dominant sense (*band[object]*) might have been influenced by two factors: how frequently the dominant sense appears in the language compared to the subordinate sense (i.e., sense dominance); or how lexically associated the verb (e.g., *get*) and the dominant sense are in the language (verb-dominant sense probability). We computed Kendall Tau partial correlation coefficients using the *ppcor* R package (Kim, 2015), and found that there is a moderate observed association between sense dominance and proportion of dominant sense choice in the lexical condition, when controlling for verb-dominant sense probability (*Tau* = .32), and a stronger observed association between verb-dominant sense probability and proportion of dominant sense choice in the lexical condition (*Tau*= .43), when controlling for sense dominance[2].

In Figure 6, we also show preliminary descriptive statistics of our current sample of 4-year-olds (*N*=11). Overall, children's pattern of responses seems qualitatively similar to adults', showing sensitivity to both lexical associations and verb-event structure. Children chose the subordinate target sense more when the story context pointed toward it (58% subordinate sense choice in control condition), while favoring the dominant sense in lexical and semantic conditions (62% dominant sense choice in lexical condition, 72% dominant sense choice in semantic condition). Statistical analyses carried out on the final data sample will allow us to test the significance of this pattern of responses.

5.5 Discussion

In sum, in this study we aim to disentangle the role that low-level lexical associations and high-level verb-event structure might have in child word sense disambiguation. The study will allow us to compare key theoretical accounts which assume different contributions of these factors early in development.

We have shown that adults can use both verb-object associations and verb-event structure to resolve lexical ambiguities. Preliminary results from children also suggest that they might be able to use both low-level (verb-object) associations and high-level (verb-event structure) cues in word sense disambiguation. If such results are confirmed, they might constitute first evidence that young children can use sentence structural cues to resolve lexical ambiguities. If children were insensitive to verb structural cues, one would expect to see no difference between performance in the control and semantic conditions (i.e., children cannot use the semantic structure of the verb *twist* and generalize it to a new object to infer that one can only twist an elastic band, not a music band).

One limitation of the current study is that, although we did our best to choose verbs that are semantically *neutral* in the lexical condition (i.e., they accept both dominant and subordinate target meanings), some differences in verb-event structure might still be present (e.g., *getting an elastic band* might still be more plausible than *getting a music band*). Nevertheless, the correlation we found between adult performance in the lexical condition and verb-dominant sense probability reassures us that frequency

[2]Although note that sense dominance and verb-dominant sense probability are based on the child-directed speech corpus and might be different in adult-directed language.

information played a prominent role in this condition despite some potential residual activation of verb-event structure.

Importantly, given the ubiquitous role of lexical associations early in development (Ambridge et al., 2015), these preliminary results highlight the importance of leveraging corpus statistics from naturalistic conversations to carefully construct experimental materials and control for the effect of lexical associations.

In the following section, we explain how we will exploit our corpus ChiSense-12 to model learning mechanisms that might underlie adult and child performance in the aforementioned experimental task.

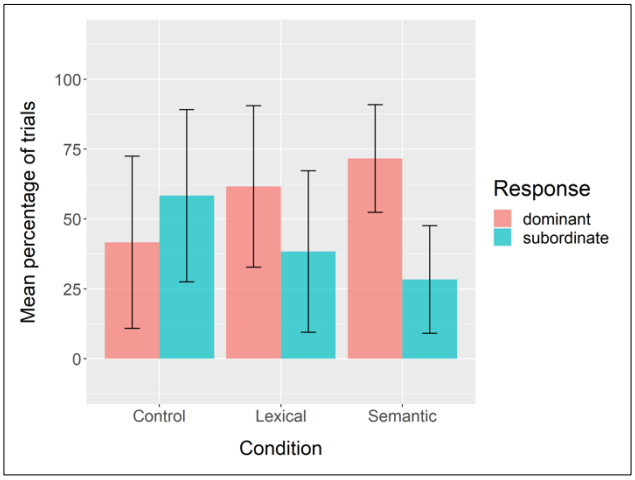

Figure 6: Mean percentage of trials in which a child (*N*=11) selected dominant or subordinate target sense, in each condition (control, lexical, semantic). Error bars show 95% confidence intervals corrected for within-subject variance.

6. Examining learning mechanisms

Our sense-tagged corpus will allow us to use a computational approach to study the learning mechanisms that might be involved in child word sense disambiguation. This computational study will have broader implications for language acquisition research, as our interest concerns domain-general learning mechanisms that under a usage-based approach to language development are thought to drive learning at multiple linguistic levels (from phonetics to pragmatics) (Bybee, 2010). Usage-based accounts assume that at least two mechanisms are involved in language learning, a simple associative-learning mechanism which is sensitive to different sources of statistical regularities in the linguistic input, and an analogical mechanism which allows the individual to carry out an analysis of common features between similar linguistic exemplars encountered and ultimately abstract linguistic structures (e.g., Abbot-Smith and Tomasello, 2006; Ambridge, 2019; 2020).

Computational implementations of these processes would require architectures which store a large amount of context-dependent information (language exemplars), while gradually being able to encode context-independent information at multiple levels of abstraction (sentence structures). Such a large problem space has often constrained model formulations to high-level descriptions or applications to a limited set of artificial tasks (e.g., Alishahi and Stevenson, 2010, Perfors et al., 2010). However, recent advances in Natural Language Processing have opened the possibility of testing usage-based models within Transformer-based architectures that possess a degree of neurobiological plausibility and can be applied to a multitude of language tasks. Although the neuropsychological realism of all low-level restrictions and biases implemented within Transformer-based models is unclear, these models are nevertheless useful to test ideas of how humans process language (e.g., Ororbia et al., 2019). For example, these models are generic neural architectures that work without implementing specific linguistic universals, being therefore in line with a usage-based view that sees the learner as able to gradually bootstrap linguistic knowledge from the input alone (e.g., Giulianelli et al., 2020; Huebner et al., 2021). In fact, usage-based models resonate closely with the learning mechanism underlying contextualized models such as BERT (Devlin et al., 2018), in which grammatical abstractions are formed at multiple levels (e.g., Clark et al., 2019; Goldberg, 2019; Hewitt and Manning, 2019; Manning et al., 2020) while retaining a large amount of context-dependent information. The generic nature of contextualized representations has also allowed researchers to model cross-modal learning (Lu et al., 2019; Qi et al., 2020; Sun et al., 2019), in line with ideas regarding domain-general applicability of mechanisms assumed by usage-based approaches.

Moreover, different versions of BERT exist, including a recent one pre-trained on child-directed speech (e.g., BabyBERTa; Huebner et al., 2021), which opens the possibility for researchers to examine a more developmentally plausible version of such models (based on realistic input children receive).

We will build on recent work evaluating Transformer-based models of word sense disambiguation (Loureiro et al., 2021), and use our sense-annotated corpus for supervised training. More specifically, Loureiro et al's framework allows to evaluate the extent to which a model trained on a sense-tagged corpus can distinguish between alternative senses of a target ambiguous word. This can be done using a feature extraction method where sense embeddings are first pre-computed by averaging the contextualized embeddings of each sense's training instances. Then, embeddings of test instances (in our case represented by the stimuli in our experiment) can be compared to pre-computed sense embeddings (e.g., via cosine similarity) to assess the degree to which a model selects a target sense in different experimental conditions. One could then assess whether cosine similarities predict the likelihood of adult and child target sense choice in our experiment (see section 5).

In sum, our project will be a unique opportunity to test core aspects of a usage-based theory by exploiting recent developments in Natural Language Processing.

7. Conclusion

We have presented the first large-scale sense-annotated corpus of child-directed speech which also allows to study the contribution of verb-event structure in early word sense disambiguation. We hope that our corpus will allow researchers to answer different questions regarding children's word sense disambiguation.

We have explained how our corpus can be used within an experimental approach to study the role of environmental factors in word sense disambiguation. Results from adults suggest that using sense-annotated corpora is important to disentangle the effect of statistical and structural sentence cues. Preliminary results from children suggest that 4-year-olds might already be able to abstract verb structural information to process ambiguous words.

Finally, we have outlined how a mixed (experimental and computational) approach can be used to study learning mechanisms in word sense disambiguation, ultimately tackling some key questions at a theoretical level.

8. Acknowledgements

The research reported in this article was supported by a Cardiff University School of Psychology PhD Studentship, and a British Academy Small Grant SRG1920\100600.

We would like to thank Jose Camacho-Collados for his comments on an earlier draft of this paper, Kelsey Frewin for her help with the creation of the stimuli for the experimental task, and Chara Sofocleous for her help as a second independent annotator for the small inter-annotator agreement study.

9. Bibliographical References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*(3).

Alishahi, A., & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, *25*(1), 50–93.

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition: *First Language*. *40*(5-6), 509-559.

Ambridge, B. (2020). Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators. *First Language*, *40*(5–6), 640–659.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2014). Child language acquisition: Why universal grammar doesn't help. *Language*, *90*(3), e53–e90.

Andreu, L., Sanz-Torrent, M., & Trueswell, J. C. (2013). Anticipatory sentence processing in children with specific language impairment: Evidence from eye movements during listening. *Applied Psycholinguistics*, *34*(1), 5–44.

Braginsky, M., Sanchez, A., & Yurovsky, D. (2019). *childesr: Accessing the 'CHILDES' Database*. R package version 0.2.1.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 276–286). Association for Computational Linguistics. arXiv:1906.04341

Connell, L., & Keane, M. T. (2004). What plausibly affects plausibility? Concept coherence and distributional word coherence as factors influencing plausibility judgments. *Memory & Cognition*, *32*(2), 185–197.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Fellbaum, C. (1998). *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Fenson, L., et al. (2007). *MacArthur-Bates communicative development inventories.* Paul H. Brookes Publishing Company Baltimore, MD.

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing Lexical Semantic Change with Contextualised Word Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973.

Goldberg, Y. (2019). *Assessing BERT's Syntactic Abilities*. arXiv preprint arXiv:1901.05287.

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid Linguistic Ambiguity Resolution in Young Children with Autism Spectrum Disorder: Eye Tracking Evidence for the Limits of Weak Central Coherence. *Autism Research*, *8*(6), 717–726.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138.

Huebner, P. A., Sulem, E., Fisher, C., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. *The SIGNLL Conference on Computational Natural Language Learning 2021*.

Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution: *Quarterly Journal of Experimental Psychology 63*(1), 160-193.

Kidd, E., & Bavin, E. L. (2005). Lexical and referential cues to sentence interpretation: An investigation of children's interpretations of ambiguous sentences. *Journal of Child Language*, *32*(4), 855-876.

Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for Statistical Applications and Methods*, *22*(6), 665–674.

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, *47*(2), 387–443.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems*, *32*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Transcription format and programs* (Vol. 1). Psychology Press.

Mani, N., Daum, M. M., & Huettig, F. (2016). "Proactive" in many ways: Developmental evidence for a dynamic pluralistic approach to prediction. *Quarterly Journal of Experimental Psychology*, *69*(11), 2189-2201.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* mit Press.

Meylan, S. C., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). Quantifying Lexical Ambiguity in Speech To and From English-Learning Children. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 252–364). Boston, MA: Cascadilla Press.

Monaghan, P., & Rowland, C. F. (2017). Combining language corpora with experimental and computational approaches for language acquisition research. *Language Learning*, *67*(S1), 14-39.

Ororbia, A. G., Mali, A., Kelly, M. A., & Reitter, D. (2019). Like a Baby: Visually Situated Neural Language Acquisition. *ArXiv:1805.11546*

Pasini, T., & Camacho-Collados, J. (2020). *A Short Survey on Sense-Annotated Corpora.* ArXiv:1802.04744.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607–642.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., & Sacheti, A. (2020). ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *ArXiv:2001.07966 [Cs]*.

Rabagliati, H., Pylkkanen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*(6), 1076–1089.

Saxton, M. (2009). The Inevitability of Child Directed Speech. In S. Foster-Cohen (Ed.), *Language Acquisition* (pp. 62–86). Palgrave Macmillan UK.

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language*, *58*, 574–608.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A Joint Model for Video and Language Representation Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7463–7472.

Trueswell, J. C., & Gleitman, L. R. (2007). Learning to parse and its implications for language acquisition. In M. G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 635–657). Oxford, United Kingdom: Oxford University Press.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1), 126–156.

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152.