



**Universiteit
Leiden**
The Netherlands

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

Dijk, B.M.A. van

Citation

Dijk, B. M. A. van. (2025, January 17). *Theory of mind in language, minds, and machines: a multidisciplinary approach*. Retrieved from <https://hdl.handle.net/1887/4176419>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4176419>

Note: To cite this publication please use the final published version (if applicable).

Bibliography

- Abdou, M., Kulmizev, A., Hershovich, D., Frank, S., Pavlick, E., and Søgaard, A. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., et al. (2024). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Adricula, N. and Narasimhan, B. (2009). ‘Understanding is Understanding by Seeing’: Visual Perception Verbs in Child Language. In *Proceedings of the 44th Boston University Conference on Language Development*, pages 18–27, Boston. Somerville, MA: Cascadilla Press.
- Agüera y Arcas, B. (2022a). Artificial neural networks are making strides towards consciousness. <https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas>. *The Economist* (Jun 11th). Accessed on: 2023-01-28.
- Agüera y Arcas, B. (2022b). Do Large Language Models Understand Us? *Daedalus*, 151(2):183–197.
- Akoglu, H. (2018). User’s guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.
- Alabdulkarim, A., Li, S., and Peng, X. (2021). Automatic story generation: Challenges and attempts. In Akoury, N., Brahman, F., Chaturvedi, S., Clark, E., Iyyer, M., and Martin, L. J., editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics.
- Alexander Pan, B. (2011). Assessing vocabulary skills. In Hoff, E., editor, *Research Methods in Child Language*, pages 100–112. John Wiley and Sons.

Bibliography

- Altszyler, E., Ribeiro, S., Sigman, M., and Fernández Slezak, D. (2017). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. In *XVIII Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 46 (Córdoba, 2017)*.
- Andreas, J. (2022). Language models as agent models. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., et al. (2023). PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403v3*.
- Apperly, I. (2012). *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press.
- Apperly, I. A., Samson, D., and Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental psychology*, 45(1):190.
- Arslan, B., Taatgen, N. A., and Verbrugge, R. (2017). Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in psychology*, 8:275.
- Astington, J. W. and Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental psychology*, 35(5):1311.
- Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.
- Baker, C., Saxe, R., and Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 33.
- Banfield, A. (1973). Narrative style and the grammar of direct and indirect speech. *Foundations of language*, 10(1):1–39.
- Barak, L., Fazly, A., and Stevenson, S. (2012). Modeling the Acquisition of Mental State Verbs. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 1–10, Montréal, Canada. Association for Computational Linguistics.
- Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme*, 34(1):74–183.
- Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

- Barone, P., Corradi, G., and Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, 57:101350.
- Beauchamp, M. H. (2017). Neuropsychology's social landscape: Common ground with social neuroscience. *Neuropsychology*, 31(8):981–1002.
- Beaudoin, C., Leblanc, E., Gagner, C., and Beauchamp, M. H. (2020). Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10.
- Beekhuizen, B., Armstrong, B. C., and Stevenson, S. (2021). Probing Lexical Ambiguity: Word Vectors Encode Number and Relatedness of Senses. *Cognitive Science*, 45(5):e12943.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Bergen, D. (2002). The role of pretend play in children's cognitive development. *Early Childhood Research & Practice*, 4(1).
- Berger, J. and Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist*, 77(4):525.
- Bian, N., Han, X., Lin, H., Lu, Y., He, B., and Sun, L. (2024). Rule or Story, Which is a Better Commonsense Expression for Talking with Large Language Models? *arXiv preprint arXiv:2402.14355*.
- Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Binz, M. and Schulz, E. (2024). Turning large language models into cognitive models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Bishop, M. J. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11:2603.
- Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. (2020). Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Bibliography

- Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, 27(11):987–989.
- Bloom, P. and German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1):B25–B31.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.
- Boogaart, R. (1999). *Aspect and temporal ordering. A contrastive analysis of Dutch and English*. PhD thesis, Vrije Universiteit Amsterdam.
- Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *arXiv preprint arXiv:2302.03494*.
- Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child language teaching and therapy*, 18(1):1–21.
- Botvin, G. J. and Sutton-Smith, B. (1977). The development of structural complexity in children’s fantasy narratives. *Developmental Psychology*, 13(4):377.
- Bowman, S. R. (2022). The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.
- Bowman, S. R. (2023). Eight Things to Know about Large Language Models. *arXiv preprint arXiv:2304.00612*.
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.
- Brahman, F., Huang, M., Tafjord, O., Zhao, C., Sachan, M., and Chaturvedi, S. (2021). “let your characters tell their story”: A dataset for character-centric narrative understanding. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen,

- M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Browning, J. and LeCun, Y. (2022). AI and The Limits Of Language. <https://www.noemamag.com/ai-and-the-limits-of-language>. *Noema Magazine* (Aug 23rd). Berggruen Institute. Accessed on: 2023-01-28.
- Bruner, J. (1990). Culture and human development: A new look. *Human development*, 33(6):344–355.
- Brunet-Gouet, E., Vidal, N., and Roux, P. (2023). Can a Conversational Agent Pass Theory-of-Mind Tasks? A Case Study of ChatGPT with the Hinting, False Beliefs, and Strange Stories Paradigms. In *International Conference on Human and Artificial Rationalities*, pages 107–126. Springer.
- Brunner, A. (2013). Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and linguistic computing*, 28(4):563–575.
- Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at Work. *Working Paper Series*, (31161). Cambridge, MA: National Bureau of Economic Research.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.
- Bunzeck, B. and Zarrieß, S. (2023). GPT-wee: How Small Can a Small Language Model Really Get? In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.
- Cancho, R. F. I. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.
- Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2):141–172.
- Cerullo, M. (2022). In Defense of Blake Lemoine and the Possibility of Machine Sentience in Lamda. <https://philarchive.org/rec/CERIDO>. *PhilPapers*. Accessed on: 2023-02-20.
- Cevoli, B., Watkins, C., Gao, Y., and Rastle, K. (2023). Shades of meaning: Uncovering the geometry of ambiguous word representations through contextualised language models. *arXiv preprint arXiv:2304.13597*.
- Chang, T. A. and Bergen, B. K. (2022). Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Bibliography

- Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., and Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More than 65 Million Words. *Psychological Science*, 32(2):218–240.
- Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage Publications.
- Chen, B. X. (2023). How ChatGPT and Bard Performed as My Executive Assistants. <https://www.nytimes.com/2023/03/29/technology/personaltech/ai-chatgpt-google-bard-assistant.html>. *New York Times* (May 5th). Accessed on: 2023-10-15.
- Chen, L., Zaharia, M., and Zou, J. (2024). How Is ChatGPT’s Behavior Changing Over Time? *Harvard Data Science Review*, (2).
- Chomsky, N. (2023). The False Promise of Chat-GPT. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>. *New York Times* (June 6th). Accessed on: 2023-01-30.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, et al. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.
- Churchland, P. M. (1985). The ontological status of observables: in praise of the superempirical virtues. *Images of science*, pages 35–47.
- Clark, H. H. (1996). *Using Language*. Cambridge University Press.
- Clement, M. (1991). Present—preterite: Tense and narrative point of view. *Linguistics in the Netherlands*, 8(1):11–20.
- Collins, K. M., Wong, C., Feng, J., Wei, M., and Tenenbaum, J. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

- Contreras Kallens, P., Kristensen-McLachlan, R. D., and Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.
- Cremin, T., Flewitt, R., Mardell, B., and Swann, J. (2017). *Storytelling in Early Childhood. Enriching language, literacy and classroom culture*. Routledge Abingdon.
- Creswell, A., Shanahan, M., and Higgins, I. (2023). Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.
- Cucchiari, C., Driesen, J., Sanders, E., et al. (2008). Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.
- Cucchiari, C. and Van hamme, H. (2013). *The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People*, pages 43–59. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Curenton, S. M. and Justice, L. M. (2008). Children’s preliteracy skills: Influence of mothers’ education and beliefs about shared-reading interactions. *Early Education and Development*, 19(2):261–283.
- Cuzzolin, F., Morelli, A., Cirstea, B., and Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in AI. *Psychological medicine*, 50(7):1057–1061.
- Dancygier, B. (2011). *The language of stories: A cognitive approach*. Cambridge University Press.
- Dancygier, B., Lu, W., and Verhagen, A. (2016). *Viewpoint and the Fabric of Meaning: Form and Use of Viewpoint Tools across Languages and Modalities*. De Gruyter Mouton.
- Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.
- David-Barrett, T. and Dunbar, R. I. M. (2016). Language as a coordination tool evolves slowly. *R. Soc. open sci.*, 3:160259.
- Davis, E. E. (2020). *Does seeing mean believing? The development of children’s semantic representations for perception verbs*. PhD thesis, The Johns Hopkins University.
- Davis, E. E. and Landau, B. (2021). Seeing and believing: the relationship between perception and mental verbs in acquisition. *Language Learning and Development*, 17(1):26–47.
- De Mulder, H. N. (2011). *Putting the pieces together: The development of theory of mind and (mental) language*. PhD thesis, Utrecht University.

Bibliography

- De Palma, P., Garcia-Camargo, L. A., Kilfoyle, J., Vandam, M., and Stover, J. (2021). Speech tested for zipfian fit using rigorous statistical techniques. *Proceedings of the Linguistic Society of America*, 6(1):394–402.
- De Villiers, J. G. (2000). Language and theory of mind: What are the developmental relationships? In Baron-Cohen, S., Tager-Flusberg, H., and Cohen, D. J., editors, *Understanding other minds: Perspectives from developmental cognitive neuroscience*, pages 83–123. Oxford University Press.
- De Villiers, J. G. (2005). Can Language Acquisition Give Children a Point of View. In Astington, J., editor, *Why language matters for theory of mind*, pages 186–219. Oxford University Press.
- De Villiers, J. G. (2007). The interface of language and Theory of Mind. *Lingua: an International Review of General Linguistics.*, 117(11):1858–1878.
- De Villiers, J. G. and De Villiers, P. A. (2014). The role of language in theory of mind development. *Topics in Language Disorders*, 34(4):313–328.
- De Vries, W., Van Cranenburgh, A., Bisazza, A., Caselli, T., Van Noord, G., and Nisim, M. (2019). BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.
- DeBuse, M. A. and Warnick, S. (2024). Plot extraction and the visualization of narrative flow. *Natural Language Engineering*, 30(3):480–524.
- Dell’Orletta, F., Montemagni, S., and Venturi, G. (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.
- Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.
- Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. (2024). Investigating Data Contamination in Modern Benchmarks for Large Language Models. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.
- Dennett, D. (1989). *The Intentional Stance*. MIT Press.
- Dennett, D. C. (1991). Two Contrasts: Folk Craft vs. Folk Science and Belief vs. Opinion. In Greenwood, J. D., editor, *The Future of Folk Psychology: Intentionality and Cognitive Science*, pages 135–48. Cambridge University Press.
- Devine, R. T. and Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child development*, 84(3):989–1003.

- Devine, R. T., Kovatchev, V., Grumley Traynor, I., Smith, P., and Lee, M. (2023). Machine learning and deep learning systems for automated measurement of “advanced” theory of mind: Reliability and validity in children and adolescents. *Psychological Assessment*, 35(2):165.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.
- Dirkson, A., Verberne, S., Van Oortmerssen, G., Gelderblom, H., and Kraaij, W. (2023). How do others cope? extracting coping strategies for adverse drug events from social media. *Journal of Biomedical Informatics*, 139:104228.
- Dockrell, J. E. and Marshall, C. R. (2015). Measurement issues: Assessing language skills in young children. *Child and Adolescent Mental Health*, 20(2):116–125.
- Dor, D. (2015). *The Instruction of Imagination. Language as a Social Communication Technology*. Oxford University Press.
- Duinmeijer, I., de Jong, J., and Scheper, A. (2012). Narrative abilities, memory and attention in children with a specific language impairment. *International Journal of Language & Communication Disorders*, 47(5):542–555.
- Dunn, L. M. and Dunn, L. M. (1997). *PPVT-III: Peabody Picture Vocabulary Test*. Circle Pines, MN: American Guidance Services.
- Ebert, K. D. and Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools*, 45(4):337–350.
- Eekhof, L. S. (2024). *Reading the Mind. The Relationship Between Social Cognition and Narrative Processing*. PhD thesis, Radboud University Nijmegen Nijmegen.
- Eekhof, L. S., Van Krieken, K., and Sanders, J. (2020). VPIP: A lexical identification procedure for perceptual, cognitive, and emotional viewpoint in narrative discourse. *Open Library of Humanities*, 6(1).
- Eldan, R. and Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *arXiv preprint arXiv:2305.07759*.
- Elias, J. (2023). Google’s newest A.I. model uses nearly five times more text data for training than its predecessor. <https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html>. CNBC (May 16th). Accessed on: 2023-05-30.

Bibliography

- Elli, G. V., Bedny, M., and Landau, B. (2021). How does a blind person see? Developmental change in applying visual verbs to agents with disabilities. *Cognition*, 212:104683.
- Enfield, N. (2023). Linguistic concepts are self-generating choice architectures. *Philosophical Transactions of the Royal Society B*, 378(1870).
- Ensink, K. and Mayes, L. C. (2010). The development of mentalisation in children from a theory of mind perspective. *Psychoanalytic Inquiry*, 30(4):301–337.
- Evanson, L., Lakretz, Y., and King, J. R. (2023). Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.
- Fengxiang, F., Yang, Y., and Yaqin, W. (2016). The probability distribution of textual vocabulary in the English language. *Journal of Quantitative Linguistics*, 23(1):49–70.
- Fernández, C. (2013). Mindful storytellers: Emerging pragmatics and theory of mind development. *First Language*, 33(1):20–46.
- Ferrer i Cancho, R. (2005). The variation of Zipf’s law in human language. *The European Physical Journal B-Condensed Matter and Complex Systems*, 44(2):249–257.
- Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. (2008). Children’s application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17:417–442.
- Floridi, L. (2023). AI as Agency without Intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1):15.
- Fludernik, M. (1996). *Towards a ‘natural’ narratology*. Routledge.
- Fodor, J. (1992). A theory of the child’s theory of mind. *Cognition*, 44(3):283–296.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT press.
- Frank, M. C. (2023). Large language models as models of human cognition. *PsyArXiv preprint psyarXiv:wxt69*.
- Frizelle, P., Thompson, P. A., McDonald, D., and Bishop, D. V. (2018). Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, 45(5):1174–1197.
- Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

- Gallagher, S. and Hutto, D. D. (2008). Understanding Others Through Primary Interaction and Narrative Practice. In Zlatev, J., Racine, T., Sinha, C., and Itkonen, E., editors, *The Shared Mind: Perspectives on Intersubjectivity*, pages 17–38. John Benjamins.
- Ganti, A., Wilson, S., Ma, Z., Zhao, X., and Ma, R. (2022). Narrative detection and feature analysis in online health communities. In Clark, E., Brahman, F., and Iyyer, M., editors, *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65, Seattle, United States. Association for Computational Linguistics.
- Gari Soler, A. and Apidianaki, M. (2021). Let’s play mono-poly: BERT can reveal words’ polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- Gildea, D. and Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.
- Grand, G., Blank, I. A., Pereira, F., and Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.
- Grant, E., Nematzadeh, A., and Griffiths, T. L. (2017). How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 39, pages 472–432.
- Grice, P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and semantics. Vol. 3: Speech acts*, pages 41–58. Academic Press, New York.
- Griffin, P., Burns, M. S., and Snow, C. E. (1998). *Preventing reading difficulties in young children*. National Academies Press.
- Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive science*, 29(2):261–290.
- Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., and Song, D. (2023). The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.
- Gurney, N., Marsella, S., Ustun, V., and Pynadath, D. V. (2021). Operationalizing theories of theory of mind: a survey. In *AAAI Fall Symposium*, pages 3–20. Springer.
- Guzman, A. L. and Lewis, S. C. (2020). Artificial intelligence and communication: A human-machine communication research agenda. *New Media & Society*, 22(1):70–86.

Bibliography

- Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv preprint arXiv:2303.13988*.
- Hagendorff, T., Fabi, S., and Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.
- Hale, C. M. and Tager-Flusberg, H. (2003). The Influence of Language on Theory of Mind: A Training Study. *Developmental science*, 6(3):346–359.
- Hansen, H. and Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3. *arXiv preprint arXiv:2202.03753*.
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters’ thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.
- Harmssen, W. N., Cucchiari, C., and Strik, H. (2021). Automatic Detection and Annotation of Spelling Errors and Orthographic Properties in the Dutch BasiScript Corpus. *Computational Linguistics in the Netherlands Journal*, 11:281–306.
- Haselton, M. G., Nettle, D., and Andrews, P. W. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 724–746.
- Heath, S. B. (1986). Taking a cross-cultural look at narratives. *Topics in language disorders*, 7(1):84.
- Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.
- Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Burstein, J., Doran, C., and Solorio, T., editors, *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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Heyes, C. (2014). False belief in infancy: a fresh look. *Developmental Science*, 17(5):647–659.
- Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, 344(6190):1243091.
- Hoeksema, J., de Glopper, K., and Van Noord, G. (2022). Syntactic profiles in secondary school writing using PaQu and SPOD. In Fišer, D. and Witt, A., editors, *CLARIN: The Infrastructure for Language Resources*. Berlin: De Gruyter.
- Hoff, E. (2006). How social contexts support and shape language development. *Developmental review*, 26(1):55–88.

- Honnibal, M. and Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- Horstmann, J. (2020). Undogmatic Literary Annotation with CATMA. In Nantke, J. and Schlupkothén, F., editors, *Annotations in Scholarly Editions and Research*, pages 157–176. De Gruyter Berlin and Boston.
- Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv preprint bioRxiv:2022.10.04.510681*.
- Hu, J. and Frank, M. (2024). Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.
- Hu, J. and Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.
- Hughes, C. and Leekam, S. (2004). What are the Links Between Theory of Mind and Social Relations? Review, Reflections and New Directions for Studies of Typical and Atypical Development. *Social development*, 13(4):590–619.
- Hutto, D. D. (2008). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. The MIT Press.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.
- Jawahar, G., Sagot, B., and Seddah, D. (2019). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.
- Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Johnson, C. (1999). Metaphor vs. Conflation in the Acquisition of Polysemy: The Case of SEE. In Hiraga, M. K., Wilcox, S., and Sinha, C., editors, *Cultural, Psychological and Typological Issues in Cognitive Linguistics: Selected papers of the bi-annual ICLA meeting in Albuquerque*, pages 155–169.
- Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Stanford University. Online resource, 3rd edition.

Bibliography

- Kadavath, S., Conerly, T., Askill, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., et al. (2022). Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.
- Kaland, N., Møller-Nielsen, A., Smith, L., Mortensen, E. L., Callesen, K., and Gottlieb, D. (2005). The Strange Stories test - a replication study of children and adolescents with Asperger syndrome. *European child & adolescent psychiatry*, 14(2):73–82.
- Karlsen, J., Hjetland, H. N., Hagtvet, B. E., Braeken, J., and Melby-Lervåg, M. (2021). The concurrent and longitudinal relationship between narrative skills and other language skills in children. *First Language*, 41(5):555–572.
- Karsdorp, F., Kranenburg, P. V., Meder, T., and Van Den Bosch, A. (2012). Casting a spell: Identification and ranking of actors in folktales. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pages 39–50. Lisbon, Portugal: Edições Colibri.
- Kidd, D., Ongis, M., and Castano, E. (2016). On literary fiction and its effects on theory of mind. *Scientific Study of Literature*, 6(1):42–58.
- Kinderman, P., Dunbar, R., and Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, (2):191–204.
- King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5):580–602.
- Kjell, O. N., Kjell, K., and Schwartz, H. A. (2023). Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Research*, page 115667.
- Knight, W. (2023). A New Chip Cluster Will Make Massive AI Models Possible. <https://www.wired.com/story/cerebras-chip-cluster-neural-networks-ai/>. *WIRED magazine* (Aug 24th). Accessed on: 2023-05-30.
- Köder, F. M. (2016). *Between direct and indirect speech: The acquisition of pronouns in reported speech*. PhD thesis, University of Groningen.
- Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.
- Kouwenhoven, T., Verhoef, T., De Kleijn, R., and Raaijmakers, S. (2022). Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers in Artificial Intelligence*, 5:886349.
- Kovatchev, V., Smith, P., Lee, M., Grumley Traynor, I., Luque Aguilera, I., and Devine, R. (2020). “What is on your mind?” Automated Scoring of Mindreading in Childhood and Early Adolescence. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6217–6228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Kuijper, S. J. M., Hartman, C. A., and Hendriks, P. (2015). Who Is He? Children with ASD and ADHD Take the Listener into Account in Their Production of Ambiguous Pronouns. *PLoS one*, 10(7):e0132408.
- Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. PhD thesis, Georgia State University.
- Laban, P., Dai, L., Bandarkar, L., and Hearst, M. A. (2021). Can Transformer Models Measure Coherence In Text: Re-Thinking the Shuffle Test. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.
- Labov, W. and Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In Holm, J., editor, *Essays on the Verbal and Visual Arts*, pages 12–44. University of Washington Press.
- Landau, B. and Gleitman, L. R. (2009). *Language and experience: Evidence from the blind child*. Harvard University Press.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.
- Lappin, S. (2023). Assessing the Strengths and Weaknesses of Large Language Models. *Journal of Logic, Language and Information*, pages 1–12.
- Laverghetta Jr, A. and Licato, J. (2021). Modeling age of acquisition norms using transformer networks. In *The International FLAIRS Conference Proceedings*, volume 34.
- Lavi-Rotbain, O. and Arnon, I. (2023). Zipfian Distributions in Child-Directed Speech. *Open Mind*, 7:1–30.
- Le, M., Boureau, Y.-L., and Nickel, M. (2019). "revisiting the evaluation of theory of mind through question answering". In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.
- Lee, Y. K., Lee, I., Park, J. E., Jung, Y., Kim, J., and Hahn, S. (2021). A Computational Approach to Measure Empathy and Theory-of-Mind from Written Texts. *arXiv preprint arXiv:2108.11810*.
- Leech, G. N. and Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. Pearson Education.
- Lenci, A. and Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.

Bibliography

- Lewis, P. A., Birch, A., Hall, A., and Dunbar, R. I. M. (2017). Higher order intentionality tasks are cognitively more demanding. *Social Cognitive and Affective Neuroscience*, 12(7):1063–1071.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2022). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations*.
- Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. (2024). The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.
- Linders, G. M. and Louwerse, M. M. (2023). Zipf’s law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review*, 30(1):77–101.
- Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Roustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
- Liveley, G. and Thomas, S. (2020). Homer’s intelligent machines. In Cave, S., Dihal, K., and Dillon, S., editors, *AI narratives. A history of imaginative thinking about intelligent machines*, pages 25–48. Oxford University Press.
- Lohmann, H. and Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, 74(4):1130–1144.
- Lorenz, T. (2023). An influencer’s AI clone will be your girlfriend for \$1 a minute. https://www.washingtonpost.com/technology/2023/05/13/caryn-ai-technology-gpt-4/?utm_campaign=wp_main&utm_medium=social&utm_source=twitter. *Washington Post* (May 13th). Accessed on: 2023-05-30.
- Lucy, L. and Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

- Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Ma, X., Gao, L., and Xu, Q. (2023a). ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 15–26, Singapore. Association for Computational Linguistics.
- Ma, Z., Pan, J., and Chai, J. (2023b). World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, Toronto, Canada. Association for Computational Linguistics.
- Ma, Z., Sansom, J., Peng, R., and Chai, J. (2023c). Towards a holistic landscape of situated theory of mind in large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.
- Maaten, L. V. D. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.
- Mani, I. (2014). Computational narratology. *Handbook of narratology*, pages 84–92.
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and 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.
- Marchetti, A., Di Dio, C., Cangelosi, A., Manzi, F., and Massaro, D. (2023). Developing ChatGPT’s theory of mind. *Frontiers in Robotics and AI*, 10:1189525.
- Marcus, G. F. (2022). Nonsense on Stilts. <https://garymarcus.substack.com/p/nonsense-on-stilts>. *Blog post* (Jun 6th). Accessed on: 2023-01-30.
- Margolin, U. (2014). *Narrator*. Hamburg University.
- Marr, B. (2023). Artificial Intimacy: How Generative AI Can Now Create Your Dream Girlfriend. <https://www.forbes.com/sites/bernardmarr/2023/09/28/artificial-intimacy-how-generative-ai-can-now-create-your-dream-girlfriend/?sh=7a1a0f72464a>. *Forbes Magazine* (May 13rd). Accessed on: 2023-10-10.

Bibliography

- McAlister, A. and Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development*, 22(2):258–270.
- McCarthy, P. M. and Jarvis, S. (2010). MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- McConachie, B. and Hart, F. E. (2006). *Performance and Cognition. Theatre Studies and the Cognitive Turn*. Routledge.
- McKeough, A. and Genereux, R. (2003). Transformation in narrative thought during adolescence: The structure and content of story compositions. *Journal of Educational Psychology*, 95(3):537.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.
- McNeill, D. (1966). The creation of language by children. In Lyons, J. and Wales, R. J., editors, *Psycholinguistics papers: The proceedings of the 1966 Edinburgh conference*. Edinburgh University Press.
- Michelmann, S., Kumar, M., Norman, K. A., and Toneva, M. (2023). Large language models can segment narrative events similarly to humans. *arXiv preprint arXiv:2301.10297*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Miller, J. F. (1991). Quantifying productive language disorders. *Research on child language disorders: A decade of progress*, pages 211–220.
- Milligan, K., Astington, J. W., and Dack, L. A. (2007). Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-Belief Understanding. *Child Development*, 78(2):622–646.
- Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in AI’s large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.
- Moghaddam, S. R. and Honey, C. J. (2023). Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.
- Monster, I., Tellings, A., Burk, W. J., Keuning, J., Segers, E., and Verhoeven, L. (2022). Word properties predicting children’s word recognition. *Scientific Studies of Reading*, 26(5):373–389.

- Nelson, K. (2005). Language Pathways into the Community of Minds. In Astington, J., editor, *Why language matters for theory of mind*, pages 26–49. Oxford University Press.
- Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., and Griffiths, T. (2018). Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.
- Nerlich, B. and Clarke, D. D. (1999). Elements for an integral theory of semantic change and semantic development. In *Meaning Change–Meaning Variation. Workshop held at Konstanz*, volume 1, pages 123–134.
- Nicolopoulou, A. (1993). Play, cognitive development, and the social world: Piaget, Vygotsky, and beyond. *Human development*, 36(1):1–23.
- Nicolopoulou, A. (2007). The interplay of play and narrative in children’s development: Theoretical reflections and concrete examples. In Goncu, A. and Gaskins, S., editors, *Play and development*, pages 249–275. Psychology Press.
- Nicolopoulou, A. (2015). Young children’s pretend play and storytelling as modes of narrative activity: From complementarity to cross-fertilization? In Douglas, S. and Stirling, L., editors, *Children’s play, pretense, and story*, pages 7–28. Routledge.
- Nicolopoulou, A. (2016). Promoting oral narrative skills in low-income preschoolers through storytelling and story acting. In Cremin, T., Flewitt, R., Mardell, B., and Swann, J., editors, *Storytelling in Early Childhood*, pages 63–80. Routledge.
- Nicolopoulou, A. (2018). Pretend and social pretend play: Complexities, continuities, and controversies of a research field. In Smith, P. K. and Roopnarine, J. L., editors, *The Cambridge Handbook of Play: Developmental and Disciplinary Perspectives*, page 183–199. Cambridge University Press.
- Nicolopoulou, A. (2019). Using a storytelling/story-acting practice to promote narrative and other decontextualized language skills in disadvantaged children. In Veneziano, E. and Nicolopoulou, A., editors, *Narrative, literacy and other skills: Studies in intervention*, pages 263–284. John Benjamins Publishing Company.
- Nicolopoulou, A., Cortina, K. S., Ilgaz, H., Cates, C. B., and de Sá, A. B. (2015). Using a narrative-and play-based activity to promote low-income preschoolers’ oral language, emergent literacy, and social competence. *Early childhood research quarterly*, 31:147–162.
- Nicolopoulou, A., Ilgaz, H., Shiro, M., and Hsin, L. B. (2022). “And they had a big, big, very long fight:” The development of evaluative language in preschoolers’ oral fictional stories told in a peer-group context. *Journal of Child Language*, 49(3):522–551.

Bibliography

- Nicolopoulou, A. and Richner, E. S. (2007). From Actors to Agents to Persons: The Development of Character Representation in Young Children’s Narratives. *Child development*, 78(2):412–429.
- Nicolopoulou, A. and Ünlütürk, B. (2017). Narrativity and Mindreading Revisited: Children’s Understanding of Theory of Mind in a Storybook and in Standard False Belief Tasks. In Ketrez, F. N., Özyürek, A., Özçalışkan, Ş., and Küntay, A. C., editors, *Social Environment and Cognition in Language Development*, pages 151–166. John Benjamins.
- Niles, J. D. (1999). *Homo narrans: The poetics and anthropology of oral literature*. University of Pennsylvania Press.
- Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666.
- Nivre, J., Zeman, D., Ginter, F., and Tyers, F. (2017). Universal Dependencies. In Klementiev, A. and Specia, L., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.
- Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, 9(4):30–33.
- Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–258.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, volume 1.
- Paape, D. (2023). When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2:202–218.
- Paley, V. G. (1990). *The Boy Who Would Be a Helicopter: The Uses of Storytelling in the Kindergarten*. Cambridge, MA: Harvard University Press.
- Parisien, C. and Stevenson, S. (2009). Modelling the acquisition of verb polysemy in children. In *Proceedings of the CogSci2009 Workshop on Distributional Semantics beyond Concrete Concepts*, pages 17–22, Austin, Texas. Cognitive Science Society.
- Patel, R. and Pavlick, E. (2022). Mapping Language Models to Grounded Conceptual Spaces. In *International Conference on Learning Representations*.

- Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.
- Pearcy, A. (2023). ‘It was as if my father were actually texting me’: grief in the age of AI. <https://www.theguardian.com/technology/2023/jul/18/ai-chatbots-grief-chatgpt>. *The Guardian* (October 15th). Accessed on: 2023-10-15.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Perner, J., Leekam, S. R., and Wimmer, H. (1987). Three-year-olds’ difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.
- Perner, J. and Wimmer, H. (1985). “John thinks that Mary thinks that...” attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.
- Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. <https://time.com/6247678/openai-chatgpt-kenya-workers/>. *Time Magazine* (Jan 18th). Accessed on: 2023-01-25.
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language Models as Knowledge Bases? In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Piantadosi, S. (2023). Modern language models refute Chomsky’s approach to language. In Gibson, E. and Poliak, M., editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414. Language Science Press.
- Piantadosi, S. and Hill, F. (2022). Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.
- Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Bibliography

- Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Propp, V. (1968). *Morphology of the Folktale*. University of Texas Press.
- Prystawski, B., Grant, E., Nematzadeh, A., Lee, S. W., Stevenson, S., and Xu, Y. (2022). The emergence of gender associations in child language development. *Cognitive Science*, 46(6):e13146.
- Quesque, F. and Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018a). Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018b). Machine theory of mind. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.
- Rabkina, I., Nakos, C., and Forbus, K. (2019). Children’s Sentential Complement Use Leads the Theory of Mind Development Period: Evidence from the CHILDES Corpus. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 2634–2639. The Cognitive Science Society.
- Regier, T., Kemp, C., and Kay, P. (2015). Word meanings across languages support efficient communication. In MacWhinney, B. and O’Grady, W., editors, *The handbook of language emergence*, pages 237–263. Wiley Online Library.
- Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.
- Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.
- Roark, B., Mitchell, M., and Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8.
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Rorty, R. (2009). *Philosophy and the Mirror of Nature*. Princeton University Press.
- Rubio-Fernández, P. (2021). Pragmatic markers: the missing link between language and Theory of Mind. *Synthese*, 199(1):1125–1158.

- Rubio-Fernández, P., Mollica, F., Oraa Ali, M., and Gibson, E. (2019). How do you know that? Automatic belief inferences in passing conversation. *Cognition*, 193:104011.
- Ryle, G. (1950). *"The Concept of Mind"*. The University of Chicago Press.
- Rączaszek-Leonardi, J., Nomikou, I., Rohlfing, K. J., and Deacon, T. W. (2018). Language development from an ecological perspective: Ecologically valid ways to abstract symbols. *Ecological Psychology*, 30(1):39–73.
- Saffran, J. R. (2020). Statistical language learning in infancy. *Child development perspectives*, 14(1):49–54.
- Sahlgren, M. and Carlsson, F. (2021). The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point. *Frontiers in Artificial Intelligence*, 4:682578.
- Samir, F., Beekhuizen, B., and Stevenson, S. (2021). A formidable ability: Detecting adjectival extremeness with DSMs. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4112–4125, Online. Association for Computational Linguistics.
- Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In Vlachos, A. and Augenstein, I., editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.
- San Juan, V. and Astington, J. W. (2017). Does language matter for implicit theory of mind? The effects of epistemic verb training on implicit and explicit false-belief understanding. *Cognitive Development*, 41:19–32.
- San Roque, L., Kendrick, K. H., Norcliffe, E., and Majid, A. (2018). Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics*, 29(3):371–406.
- San Roque, L. and Schieffelin, B. B. (2019). Perception verbs in context: Perspectives from Kaluli (Bosavi) child-caregiver interaction. In O’Meara, C., Speed, L. J., San Roque, L., and Majid, A., editors, *Perception Metaphors*, pages 347–368. John Benjamins.
- Sanders, J. (2010). Intertwined voices: Journalists’ modes of representing source information in journalistic subgenres. *English Text Construction*, 3(2):226–249.
- Santana, B., Campos, R., Amorim, E., Jorge, A., Silvano, P., and Nunes, S. (2023). A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Bibliography

- Sap, M., Le Bras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Schaerlaekens, A. M. and Gillis, S. (1993). *De taalverwerving van het kind: een hernieuwde oriëntatie in het Nederlandstalig onderzoek*. Wolter-Nordhoff.
- Schank, R. C. (1995). *Tell me a story: Narrative and intelligence*. Northwestern University Press.
- Schlangen, D. (2021). Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.
- Schlichting, J. E. P. T. (1996). *Discovering syntax: An empirical study in Dutch language acquisition*. Nijmegen: Nijmegen University Press.
- Schlinger, H. D. (2009). Theory of mind: An overview and behavioral perspective. *The Psychological Record*, 59:435–448.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.
- Sejnowski, T. (2022). Large language models and the reverse turing test. *arXiv preprint arXiv:2207.14382*.
- Sellars, W. (1956). *Empiricism and the Philosophy of Mind*. University of Minnesota Press, Minneapolis.
- Shanahan, M. (2024). Talking about large language models. *Commun. ACM*, 67(2):68–79.
- Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. (2024). Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian’s, Malta. Association for Computational Linguistics.
- Shapiro, L. R. and Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children’s picture-elicited narratives. *Developmental psychology*, 27(6):960.

- Sharma, A., Miner, A., Atkins, D., and Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, On-line. Association for Computational Linguistics.
- Shatz, M., Diesendruck, G., Martinez-Beck, I., and Akar, D. (2003). The influence of language and socioeconomic status on children’s understanding of false belief. *Developmental Psychology*, 39(4):717.
- Siegal, M. and Beattie, K. (1991). Where to look first for children’s knowledge of false beliefs. *Cognition*, 38(1):1–12.
- Sileo, D. and Lernould, A. (2023). MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore. Association for Computational Linguistics.
- Slade, L. and Ruffman, T. (2005). How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory and false belief. *British Journal of Developmental Psychology*, 23(1):117–141.
- Smith, P. K. and Roopnarine, J. L. (2018). *The Cambridge Handbook of Play: Developmental and Disciplinary Perspectives*. Cambridge University Press.
- Smith, R. D. (2007). Investigation of the Zipf-plot of the extinct Meroitic language. *Glottometrics*, 15:53–61.
- Snow, C. E. and Dickinson, D. K. (1991). *Some skills that aren’t basic in a new conception of literacy*, pages 179–191. State University of New York Press.
- Southwood, F. and Russell, A. F. (2004). Comparison of Conversation, Freeplay, and Story Generation as Methods of Language Sample Elicitation. *Journal of Speech, Language and Hearing Research*, 47(2):366–376.
- Sparrow, J. (2022). ‘Full-on robot writing’: the artificial intelligence challenge facing universities. <https://www.theguardian.com/australia-news/2022/nov/19/full-on-robot-writing-the-artificial-intelligence-challenge-facing-universities>. *The Guardian* (November 18th). Accessed on: 2023-01-25.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., et al. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Bibliography

- Stammbach, D., Antoniak, M., and Ash, E. (2022). Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In Clark, E., Brahman, F., and Iyyer, M., editors, *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.
- Stiller, J. and Dunbar, R. I. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104.
- Strik, H., Loo, J. V. D., Doremalen, J. V., and Cucchiarini, C. (2010). Practicing syntax in spoken interaction: automatic detection of syntactical errors in non-native utterances. In *Proc. Second Language Studies: Acquisition, Learning, Education and Technology (L2WS 2010)*.
- Stuart, M. T. and Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27.
- Subramonian, A., Yuan, X., Daumé III, H., and Blodgett, S. L. (2023). It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.
- Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, 153(4):1066–1075.
- Sutton-Smith, B. (1986). Children’s fiction making. In Crites, S. and Sarbin, T. R., editors, *Narrative psychology: The storied nature of human conduct*, pages 67–90. Praeger Publishers/Greenwood Publishing Group.
- Sutton-Smith, B. (2012). *The folkstories of children*. University of Pennsylvania Press.
- Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*, volume 54. Cambridge University Press.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1):24–54.
- Tellings, A., Hulsbosch, M., Vermeer, A., and Van den Bosch, A. (2014). BasiLex: An 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208.
- Tellings, A., Oostdijk, N., Monster, I., Grootjen, F., and Van Den Bosch, A. (2018a). BasiScript: A corpus of contemporary Dutch texts written by primary school children. *International Journal of Corpus Linguistics*, 23(4):494–508.

- Tellings, A., Oostdijk, N., Monster, I., Grootjen, F., and Van Den Bosch, A. (2018b). Spelling errors of 24 cohorts of children across primary school 2012-2015: a Basiscript corpus study. *Computational Linguistics in the Netherlands Journal*, 8:83–98.
- Tomasello, M. (2003). The key is social cognition. *Language in mind: Advances in the study of language and thought*, pages 44–57.
- Tomasello, M. (2008). *Origins of Human Communication*. MIT Press, Cambridge, MA.
- Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology*, 44(3):187–194.
- Tomasello, M. et al. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, 103130:103–130.
- Tompkins, V., Farrar, M. J., and Montgomery, D. E. (2019). Speaking Your Mind: Language and Narrative in Young Children’s Theory of Mind Development. *Advances in Child Development and Behavior*, 56:109–140.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.
- Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. (2023). Do Large Language Models know what humans know? *Cognitive Science*, 47(7):e13309.
- Tsfasman, M., Fenech, K., Tarvirdians, M., Lorincz, A., Jonker, C. M., and Oertel, C. (2022). Towards creating a conversational memory for long-term meeting support: predicting memorable moments in multi-party conversations through eye-gaze. In *ICMI 2022-Proceedings of the 2022 International Conference on Multimodal Interaction*. Association for Computing Machinery (ACM).
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX:433–460.
- Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- Vala, H., Jurgens, D., Piper, A., and Ruths, D. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In Márquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Bibliography

- Van Den Bosch, A., Busser, B., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected papers of the 17th computational linguistics in the Netherlands meeting*, pages 99–114. CLIN Leuven, BE.
- Van Der Meulen, A., Roerig, S., de Ruyter, D., van Lier, P., and Krabbendam, L. (2017). A comparison of children’s ability to read children’s and adults’ mental states in an adaptation of the reading the mind in the eyes task. *Frontiers in psychology*, 8:594.
- Van Duijn, M. and Verhagen, A. (2018). Beyond triadic communication: A three-dimensional conceptual space for modelling intersubjectivity. *Pragmatics & Cognition*, 25(2):384–416.
- Van Duijn, M. J. (2016). *The lazy mindreader. A humanities perspective on mindreading and multiple-order intentionality*. PhD thesis, Leiden University.
- Van Duijn, M. J., Sluiter, I., and Verhagen, A. (2015). When narrative takes over: The representation of embedded mindstates in Shakespeare’s Othello. *Language and Literature*, 24(2):148–166.
- Van Eecke, P., Verheyen, L., Willaert, T., and Beuls, K. (2023). The candide model: How narratives emerge where observations meet beliefs. In Akoury, N., Clark, E., Iyyer, M., Chaturvedi, S., Brahman, F., and Chandu, K., editors, *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 48–57, Toronto, Canada. Association for Computational Linguistics.
- Van Elk, R., Lanser, D., and Gerritsen, S. (2012). *Relatie Opleidingsniveau en Arbeidsaanbod*. Netherlands Bureau for Economic Policy Analysis (CPB).
- Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
- Van Haastrecht, M., Brinkhuis, M., Peichl, J., Remmele, B., and Spruit, M. (2023). Embracing trustworthiness and authenticity in the validation of learning analytics systems. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 552–558.
- Van Koert, M., Hollebrandse, B., and Van Hout, A. (2010). Gaan ‘go’ as dummy auxiliary in Dutch children’s tense production. *GAGL: Groninger Arbeiten zur germanistischen Linguistik*, (51):43–54.
- Van Noord, G. (2006). At Last Parsing Is Now Operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42.
- Van Schuppen, L., Van Krieken, K., and Sanders, J. (2020). Variations in Viewpoint Presentation: The ‘Pear Story’ as Told by People with a Schizophrenia Diagnosis. *Open Library of Humanities*, 6(2):2.
- Vandelanotte, L. (2009). *Speech and thought representation in English: A cognitive-functional approach*. Mouton de Gruyter.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Verhagen, A. (2005). *Constructions of Intersubjectivity. Discourse, Syntax, and Cognition*. Oxford University Press.
- Verhagen, A. (2015). Grammar and cooperative communication. In Dabrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 232–252. De Gruyter Mouton, Berlin, München, Boston.
- Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied psycholinguistics*, 22(2):217–234.
- Vermeule, B. (2009). *Why Do We Care About Literary Characters?* John Hopkins University Press.
- Verstraten, P. (2009). *Film Narratology*. University of Toronto Press.
- Voskuhl, A. (2019). *Androids in the enlightenment: Mechanics, artisans, and cultures of the self*. University of Chicago Press.
- Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing Pretrained Language Models for Lexical Semantics. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.
- Wang, Y. and Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135–147.
- Wardetzky, K. (1990). The structure and interpretation of fairy tales composed by children. *Journal of American Folklore*, pages 157–176.
- Warstadt, A. and Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In Lappin, S. and Bernardy, J.-P., editors, *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.
- Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
- Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Bibliography

- Wei, C., Wang, Y.-C., Wang, B., and Kuo, C.-C. J. (2023). An overview on language models: Recent developments and outlook. *arXiv preprint arXiv:2303.05759*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., and Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *CoRR*, abs/2201.11903.
- Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6):728–755.
- Wellman, H. M. and Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, 75(2):523–541.
- Wevers, M. (2019). Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In Tahmasebi, N., Borin, L., Jatowt, A., and Xu, Y., editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy. Association for Computational Linguistics.
- Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.
- Wijnen, F. and Verrips, M. (1998). The Acquisition of Dutch Syntax. In Wijnen, F. and Verrips, M., editors, *The Acquisition of Dutch*, pages 223–300. John Benjamins Publishing Company.
- Wilcox, E. G., Futrell, R., and Levy, R. (2023). Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.
- Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2022). Learning syntactic structures from string input. *Algebraic Structures in Natural Language*, pages 113–138.
- Wilson, R., Hruby, A., Perez-Zapata, D., Van Der Kleij, S. W., and Apperly, I. A. (2023). Is recursive “mindreading” really an exception to limitations on recursive thinking? *Journal of Experimental Psychology: General*, 152(5):1454–1468.
- Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, 13(1):103–128.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Xiao, H. (2008). On the Applicability of Zipf’s Law in Chinese Word Frequency Distribution. *Journal of Chinese Language and Computing*, 18(1):33–46.

- Yamshchikov, I. and Tikhonov, A. (2023). What is wrong with language models that can not tell a story? In Akoury, N., Clark, E., Iyyer, M., Chaturvedi, S., Brahman, F., and Chandu, K., editors, *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 58–64, Toronto, Canada. Association for Computational Linguistics.
- Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al. (2023). A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv preprint arXiv:2303.10420*.
- Yu, S., Xu, C., and Liu, H. (2018). Zipf’s law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. *CoRR*, abs/1807.01855.
- Zeman, S. (2016). Perspectivization as a link between narrative micro-and macro-structure. In Igl, N. and Zeman, S., editors, *Perspectives on Narrativity and Narrative Perspectivization*, pages 17–42. John Benjamins Amsterdam.
- Zeman, S. (2018). What is a narration—and why does it matter. In Hübl, A. and Steinbach, M., editors, *Linguistic foundations of narration in spoken and sign language*, pages 173–206. Amsterdam/Philadelphia: Benjamins.
- Zeman, S. (2020). Parameters of narrative perspectivization: The narrator. *Open Library of Humanities*, 6(2):28.
- Zhao, X., Wang, T., Osborn, S., and Rios, A. (2023). BabyStories: Can reinforcement learning teach baby language models to write better stories? In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 186–197, Singapore. Association for Computational Linguistics.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., YU, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023a). LIMA: Less Is More for Alignment. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.
- Zhou, K., Jurafsky, D., and Hashimoto, T. (2023b). Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.
- Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press.

Bibliography

Zunshine, L. (2006). *Why We Read Fiction: Theory of the Mind and the Novel*. Ohio State University Press.

Zunshine, L. (2019). What Mary Poppins knew: Theory of Mind, children's literature, history. *Narrative*, 27(1):1–29.

