Research Statement

Daniel Quigley, Linguistics Department, University of Wisconsin - Milwaukee

My dream is to build machines that can navigate without human supervision the ambiguities of human language. Achievements in machine learning (ML) and natural language processing (NLP) — areas of artificial intelligence (AI) — have facilitated interactions between humans and machines via written and spoken natural language. Contemporary ML methods for NLP rely on labeled data, which has become one of the key bottlenecks in creating systems that can attain a human-level understanding of the nuances of language. Indeed, most real-world applications of AI are limited to specific domains in which many carefully labeled data can be unambiguously and easily acquired. Since annotation is expensive and laborious, any synergies with existing NLP tasks are useful and enable us to leverage auxiliary data when learning models for complex language phenomena such as ellipsis resolution.

To address this limitation, ***my research focuses on learning, describing, and implementing the underlying linguistics of ellipsis phenomena to develop frameworks for machines such that interpretation of intra- and extra-linguistic context to resolve elliptical constructions with minimal human supervision can be achieved***. Below, I briefly describe ellipsis and relevant areas of NLP, and conclude with the *raison d'être*, which includes how my work integrates with adjacent areas of NLP and ML research.

1 Elliptical Constructions in Language

Ellipsis is a linguistic phenomenon in which some parts of sentences are left unexpressed [4]. The elided — deleted — material can be of any kind of various syntactic constituents at the clausal, predicate, and nominal levels [4, 21]. The examples in (1) are the various kinds of elliptical constructions; where appropriate, the items in angled brackets < > are the elided material.

(1) **Examples of ellipsis**

a. **NP ellipsis** involves an elision of the noun phrase constituent of a determiner phrase.
   i. *Three old cars do not cost as much as two new* <cars>.

b. **VP ellipsis** is an elision of the verb phrase/main sentence predicate (excluding any finite auxiliary).
   i. *Noam didn't write a book, but Daniel might* <write a book>.

c. **Sluicing** is an elision of an entire (embedded) clause, leaving the leftmost *wh*-word intact.
   i. *Daniel said something to Noam, but no one knows what* <she said>.

d. **Sprouting** is a subtype of sluicing where the *wh*-word has no overt correlate in the antecedent.
   i. *Daniel is coming home, but she won't tell us what time*.

e. **Swiping** is a subtype of sluicing in which the sprout has a prepositional phrase remnant.
   i. *Noam is working on it, but I'm not sure who with* <he is working on it>.

f. **Gapping** is clausal ellipsis in which a verb (and auxiliaries) is removed in a series of coordinations.
   i. *Noam wrote a book, and Daniel* <wrote> *a song*.

g. **Pseudogapping** is an ellision of the predicate (not finite auxiliaries) in a series of coordinations.
   i. *Noam wrote a book, and Daniel did* <write> *a song*.

h. **Stripping** is a coordinate clausal ellipsis, leaving behind a single (non-*wh*) remnant (cf. sluicing).
   i. *Noam read the newspaper and Daniel* <read the newspaper> *too*.

i. **Null complement anaphora** is elision of an entire predicative or nonfinite embedded constituent.
   i. *Daniel asked Noam to read a book, but he refused* <to read a book>.

j. **Comparative deletion** is an ellipsis in comparative clauses.
   i. *Noam wrote a more boring book than Daniel* <wrote> *a poem*.

Resolution of elliptical constructions can be described syntactically (defined over phrase markers or syntactic derivations) [28, 8], semantically (defined over semantic representations or computations) [13, 12], or by some treatment of both [15, 22]. Three main questions have occupied much of the literature [21], stated in (2), each with their relevant related questions: (2a) the **structure** question, (2b) the **identity** question, and (2c) the **licensing** question.

(2) **Questions surrounding ellipsis**

a. What is the structural nature of the ellipsis site?
   i. Is there internal syntactic structure in the elided site?
   ii. Is the ellipsis site a silent pro-form? Is it something else? How can we tell?

b. The understood material is identical to some antecedent; is the identity syntactic or semantic?
   i. Does the antecedent need to be explicitly linguistically mentioned?
   ii. Can it simply be something pragmatically/discursively salient?

c. What kinds of material can be elided, and what are the locality conditions on the relation between these structures and ellipsis?
   i. Under what conditions can ellipsis occur?
   ii. Why can some constituents be elided while others cannot?

2 Research Areas

Ellipsis continues to be of interest to linguists exactly because such constructions have meaning without salient form, and unresolved ellipses mask information to a machine that is otherwise available to a human participant in speech; furthermore, it is an important source of error in machine translation [20], question answering [30, 1], and dialogue understanding [3, 25]. Consequently, there is a need for NLP innovations to automatically detect and accurately interpret elliptical constructions in human speech. Key research areas implemented in my own work in this domain are summarized in the paragraphs below:

- **Anaphora resolution**: [11] were the first to treat ellipsis as a species of **anaphora**: a piece of linguistic material that gets its denotation from a salient antecedent [11, 21]. Anaphora resolution (AR) is an important area of research in NLP; it plays a substantial role in complex tasks such as information extraction, question answering, and machine translation. AR can be treated as a sequence of two separate tasks: anaphor identification (detection) selecting items as anaphors and antecedent selection (resolution) creates the link between the anaphor and the antecedent. [19] describe an NLP pipeline for resolving VP ellipsis that expands on the two main tasks of AR:

  1. **target detection**, where the subset of ellipsis targets is identified
  2. **antecedent head resolution**, where, for each detected target, identify potential antecedent heads
  3. **antecedent boundary determination**, where the model constructs boundaries for the antecedent

  Practical ellipsis algorithms for NLP that might be enhanced by methods in AR is a driving focus of my work. I first investigate how to interpret ellipsis-as-anaphor with local intra-linguistic contexts prior to working with longer distance contexts; extra-linguistic contexts follow.

- **Reformulation**: [1] recast sluicing and VP ellipsis as machine reading comprehension problems. [14] recast bridging AR as question answering (QA) based on context, and [31, 18] also reformulate coreference resolution (CR) — resolution of all mentions that refer to the same real world entity — and named entity recognition as QA. Ellipsis and questions put in focus referentially dependent expressions [2], or free variables [24], that need to be resolved in order to comprehend the discourse; this lends itself well to methods developed for AR.

  Recasting elliptical constructions into similar NLP problems is advantageous for the useful auxiliary data and computational methods they provide. I intend to investigate the usefulness of recasting elliptical constructions into similar NLP frameworks as methods to resolve ellipsis resolution.

- **Neural networks**: [32] applies a neural network model for VP ellipsis. The authors apply a **support-vector machine** (SVM) model with non-linear kernel function as a classification task for identifying ellipsis resolution over a long distance, and a **multilayer perceptron** (MLP) — a neural network with a hidden layer — and the transformer — a kind of long short-term memory (LSTM) with less computational time — as the neural models.

  [16] highlight the syntactic and semantic characteristics of ellipsis, and demonstrate robust scores through pretrained **BERT** (Bidirectional Encoder Representations from Transformers) embeddings for word representations and the importance of manual features. For the classification subtasks of ellipsis resolution, the authors outline how the detection of ellipsis follows from a simple MLP, and how a **recurrent neural network** (RNN) model is a better choice for the resolution step. [23, 29] show that improved accuracy scores result from combining neural network models in tasks of AR and CR with a **multi-pass sieve** architecture[26, 17].

  The main disadvantage of using neural networks is the clustering time, which is way longer than in compared approaches. Even considering the time-impediment, neural networks with multi-pass sieve architectures show promise in resolving reference problems in NLP. I intend to implement such methods to attempt resolution of ellipsis constructions.

- **Construction grammar**: NLP systems tend to frame linguistic structures as (more or less) modular units — phonemes, morphemes, words, syntax, discourse — and errors at one level propagate to the next. Spontaneous human speech, dialectal variation, and idiolects tend to break these abstraction barriers. Principles from **construction grammar** (CxG) offer enormous potential for resolving these problems. In the CxG framework, linguistic patterns (at any scale) are paired with meanings that combine to create utterances and their semantic representations [9, 5]. This obviates the need to define discretion between morphemes, words, and syntax, allowing multi-word expressions, idiosyncratic syntactic constructions, and productive morphology to flourish alongside the usual NLP categories [7].

  It is not necessary to rewrite the standard NLP pipeline to apply the key insights of CxG. FrameNet [27] represents **semantic frames** as indicated by **lexical unit** (LU) "targets"; as long as at least one relevant lexical or morphological span exists, the targets can be expanded without much trouble to allow richer, more flexible spans. With constructions linked directly to semantic frames, automatic taggers can rely on the usual robust NLP tags and parsers to determine the presence of a given construction and its components. This approach to NLP gains much of the representational flexibility of constructions, while still retaining the ability to use existing NLP infrastructure.

  In the longer term, my hope is that the CxG and NLP communities will work together to define more flexible representations for semantic frames. It is a novel approach to apply this framework to ellipsis resolution in NLP, informed by its CxG treatment in [6, 10].

3 Integration of Research

To summarize, my research in ellipsis resolution focuses on the underlying linguistics and developing algorithms that learn to understand them with weak or no human supervision; this integrates naturally with key areas of ML research:

- **Human language technologies** accounts for the interaction gap between humans and machines. Human language understanding relies critically on the ability to obtain unambiguous representations of linguistic content. While some ambiguities can be resolved using intra-linguistic contextual cues, the disambiguation of many linguistic constructions requires the integration of world knowledge and perceptual information obtained from other modalities.

- **Computer vision** is the premier extra-linguistic method for antecedents. Often, an antecedent is not a salient linguistic element, but some entity from the real, visual world. Enabling machines to understand what they see provides the extra-linguistic context necessary to resolve various anaphora problems. Signed languages exhibit the phenomenon as well, and operate on the same theoretical linguistic principles. Incorporating signed language is extremely important so as not to discriminate or be otherwise biased against an entire community of humans interacting with machines.

- **Human-computer interaction** strives to account for accessibility and inclusion. How can we create unbiased and secure language perception algorithms? Ellipsis, AR, and CR are cross-linguistically frequent, yet within those typologies more complicated matter must be taken into account; signed languages, dialectal variations, personal idiolects, and speech-related deviancy from some standard form should be adapted to by machines to avoid biases towards or against a speaker.

I believe that this work will lead us towards an optimistic future of rich and amazing experiences of human-machine communication. I am excited about the challenges of this interdisciplinary research, in collaboration with (but not limited to) colleagues in linguistics, cognitive science, neuroscience, mathematics, and computer science. Collaboration is not without its own challenges, but it is also the most fun and effective way to perform research; I will work to continue to be at the forefront of that research as a linguist. I am passionate about uncovering the right questions, exploring innovative and creative solutions, and communicating them to the academic, industrial, and popular communities.

References

[1] Rahul Aralikatte et al. "Ellipsis Resolution as Question Answering: An Evaluation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Association for Computational Linguistics, 2021, pp. 810–817.

[2] Gregory Carlson. "Anaphora". In: Jan. 2006. ISBN: 9780470018866. DOI: 10.1002/0470018860.s00212.

[3] Tagyoung Chung and Daniel Gildea. "Effects of Empty Categories on Machine Translation". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Association for Computational Linguistics, 2010, pp. 636–645.

[4] Jeroen van Craenenbroeck and Jason Merchant. "Ellipsis phenomena". In: *The Cambridge Handbook of Generative Syntax*. Ed. by Marcel den Dikken. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2013, pp. 701–745.

[5] William Croft. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, 2001. ISBN: 9780198299547.

[6] Peter W. Culicover and Ray Jackendoff. "Ellipsis in Simpler Syntax". In: *The Oxford Handbook of Ellipsis*. Ed. by Jeroen vanCraenenbroeck and Tanja Temmerman. Oxford University Press, 2019. ISBN: 9780198712398.

[7] Jesse Dunietz, Lori S. Levin, and Miriam R. L. Petruck. "Construction Detection in a Conventional NLP Pipeline". In: *AAAI Spring Symposia*. 2017.

[8] Robert Fiengo and Robert May. *Indices and identity*. Cambridge, Mass: MIT Press, 1994.

[9] Adele E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, 1995. ISBN: 9780226300863.

[10] Adele E. Goldberg and Florent Perek. "Ellipsis in Construction Grammar". In: *The Oxford Handbook of Ellipsis*. Ed. by Jeroen vanCraenenbroeck and Tanja Temmerman. Oxford University Press, 2019. ISBN: 9780198712398.

[11] Jorge Hankamer and Ivan Sag. "Deep and Surface Anaphora". In: *Linguistic Inquiry* 7.3 (1976), pp. 391–428.

[12] Daniel Hardt. "Dynamic Interpretation of Verb Phrase Ellipsis". In: *Linguistics and Philosophy* 22.2 (1999), pp. 185–219.

[13] Irene Heim. "Predicates or Formulas? Evidence from Ellipsis". In: *Semantics and Linguistic Theory* 7 (Jan. 1997).

[14] Yufang Hou. “Bridging Anaphora Resolution as Question Answering”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 1428–1438.

[15] Andrew Kehler. “Coherence and the Resolution of Ellipsis”. In: *Linguistics and Philosophy* 23.6 (2000), pp. 533–575.

[16] Payal Khullar. “Exploring Statistical and Neural Models for Noun Ellipsis Detection and Resolution in English”. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 139–145.

[17] Heeyoung Lee et al. “Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules”. In: *Computational Linguistics* 39.4 (Dec. 2013), pp. 885–916.

[18] Xiaoya Li et al. “A Unified MRC Framework for Named Entity Recognition”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 5849–5859.

[19] Zhengzhong Liu, Edgar Gonzàlez Pellicer, and Daniel Gillick. “Exploring the steps of Verb Phrase Ellipsis”. In: *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 32–40.

[20] Vivien Macketanz, Aljoscha Avramidis Eleftherios and Burchardt, and Hans Uszkoreit. “Fine-grained evaluation of German-English Machine Translation based on a Test Suite”. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Oct. 2018, pp. 578–587.

[21] Jason Merchant. “Ellipsis: A survey of analytical approaches”. In: *The Oxford Handbook of Ellipsis*. Ed. by Jeroen van Craenenbroeck and Tanja Temmerman. Oxford University Press, 2018. ISBN: 9780198712398.

[22] Jason Merchant. “Voice and Ellipsis”. In: *Linguistic Inquiry* 44.1 (Jan. 2013), pp. 77–108.

[23] Bartlomiej Niton, Paweł Morawiecki, and Maciej Ogrodniczuk. “Deep Neural Networks for Coreference Resolution for Polish”. In: *LREC*. 2018.

[24] Barbara H. Partee. “Bound Variables and Other Anaphors”. In: *Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*. TINLAP ’78. Urbana-Champaign, Illinois: Association for Computational Linguistics, 1978, pp. 79–85.

[25] Victor Petrén Bach Hansen and Anders Søgaard. “What Do You Mean ‘Why?’: Resolving Sluices in Conversations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 7887–7894.

[26] Karthik Raghunathan et al. “A Multi-Pass Sieve for Coreference Resolution”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 492–501.

[27] Josef Ruppenhofer et al. *FrameNet II: Extended Theory and Practice*. Distributed with the FrameNet data. Berkeley, California: International Computer Science Institute, 2006.

[28] Ivan Andrew Sag. “Deletion and Logical Form”. PhD thesis. 1976.

[29] Nikolaos Stylianou and Ioannis Vlahavas. “A neural Entity Coreference Resolution review”. In: *Expert Systems with Applications* 168 (2021).

[30] Antonio Vicedo José L. and Ferrández. “Importance of Pronominal Anaphora Resolution in Question Answering Systems”. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, Oct. 2000, 555–562".

[31] Wei Wu et al. “CorefQA: Coreference Resolution as Query-based Span Prediction”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 6953–6963.

[32] Wei-Nan Zhang et al. “A Neural Network Approach to Verb Phrase Ellipsis Resolution”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 7468–7475.