The State of Linguistic Databases: A Critical Evaluation of Current Resources

Alireza Dehbozorgi

3 min read · Just now

Listen Share ••• More

Linguistic databases are an essential tool for researchers in the field of linguistics. They provide a rich source of data that can be analyzed to gain insights into the structure and use of language. However, the quality and usefulness of these databases vary widely, and there are several challenges and limitations that must be considered when using them for research purposes. In this concise critical essay, I will evaluate the current state of linguistic databases and discuss some of the challenges and limitations associated with their use.

One of the main challenges associated with linguistic databases is the quality and accuracy of the data. Many databases rely on information that has been collected from non-experts or from sources that may not be reliable. This can lead to errors and inconsistencies in the data, which can affect the results of linguistic analyses. Some databases also suffer from a lack of diversity, with a focus on a narrow range of languages or dialects, which can limit the generalizability of the findings.

Another challenge associated with linguistic databases is the lack of standardization in data collection and annotation. There is often a lack of consistency in the way that linguistic features are annotated or labeled, which can make it difficult to compare data across different databases or studies. This can also make it difficult to replicate studies or to build on previous research.

Despite these challenges, there are several linguistic databases that are widely used and have been shown to be useful for linguistic research. For example, the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC) contain large amounts of data from a wide range of sources and have been used to study various aspects of language use and variation. The Universal Dependencies (UD) project provides a framework for annotating dependency syntax in a standardized way, which has helped to improve consistency and comparability across different languages.

However, there is still much room for improvement in the development and use of linguistic databases. One possible way to address some of the challenges and limitations is to encourage greater collaboration and standardization across different databases and studies. This could involve the development of common annotation schemes and guidelines, as well as the sharing of data and resources across different research projects and institutions.

Another potential solution is to use advances in technology, such as machine learning and natural language processing, to improve the quality and accuracy of linguistic databases. For example, automated methods could be used to annotate linguistic features or to detect errors and inconsistencies in the data.

In conclusion, linguistic databases are a valuable resource for linguistic research, but they also come with several challenges and limitations. The quality and accuracy of the data can vary widely, and there is often a lack of standardization in data collection and annotation. However, with greater collaboration and standardization, as well as advances in technology, it may be possible to address some of these challenges and improve the usefulness and reliability of linguistic databases.

. . .

References:

- Lüdeling, A. & M. Kytö (Eds.), (2008) Corpus linguistics: An international handbook (2 Vols.). Berlin: Walter de Gruyter.
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. Sebastopol, CA: O'Reilly Media.
- Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.
- Haspelmath, M. (2010). Comparative concepts and descriptive categories in cross-linguistic studies. Language, 86(3), 663–687.
- Greenberg, J. H. (1966). Language universals: With special reference to feature hierarchies. The Hague: Mouton.
- Yang, Z., Levow, G-A., & Meng, H. M. (2013). Crowdsourcing for Spoken Dialog System Evaluation. In Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, London: Wiley (pp. 217–240).
- https://www.linkedin.com/advice/3/what-advantages-disadvantages-different-3e
- Strunk, J., & White, E. (2003). The elements of style (4th ed.). New York: Longman.

- Hovy, D., & Lin, C. (1999). Automated text summarization in SUMMARIST. In D. Marcu (Ed.), Advances in automatic text summarization (pp. 141–165). Cambridge, MA: MIT Press.

. . .

Thank you so much for your precious time and attention! Stay tuned for more!

. . .

Alireza Dehbozorgi

Email: alirezadehbozorgi83@yahoo.com

Twitter

LinkedIn

Mastodon

Facebook

Instagram

Github

Linguistics Database Corpus Linguistics Data Science Naturallanguageprocessing

Edit profile

Written by Alireza Dehbozorgi

59 Followers

I'm a linguist and AI researcher interested in mathematical/computational approaches to both human and formal languages. Twitter: @BDehbozorgi83

More from Alireza Dehbozorgi

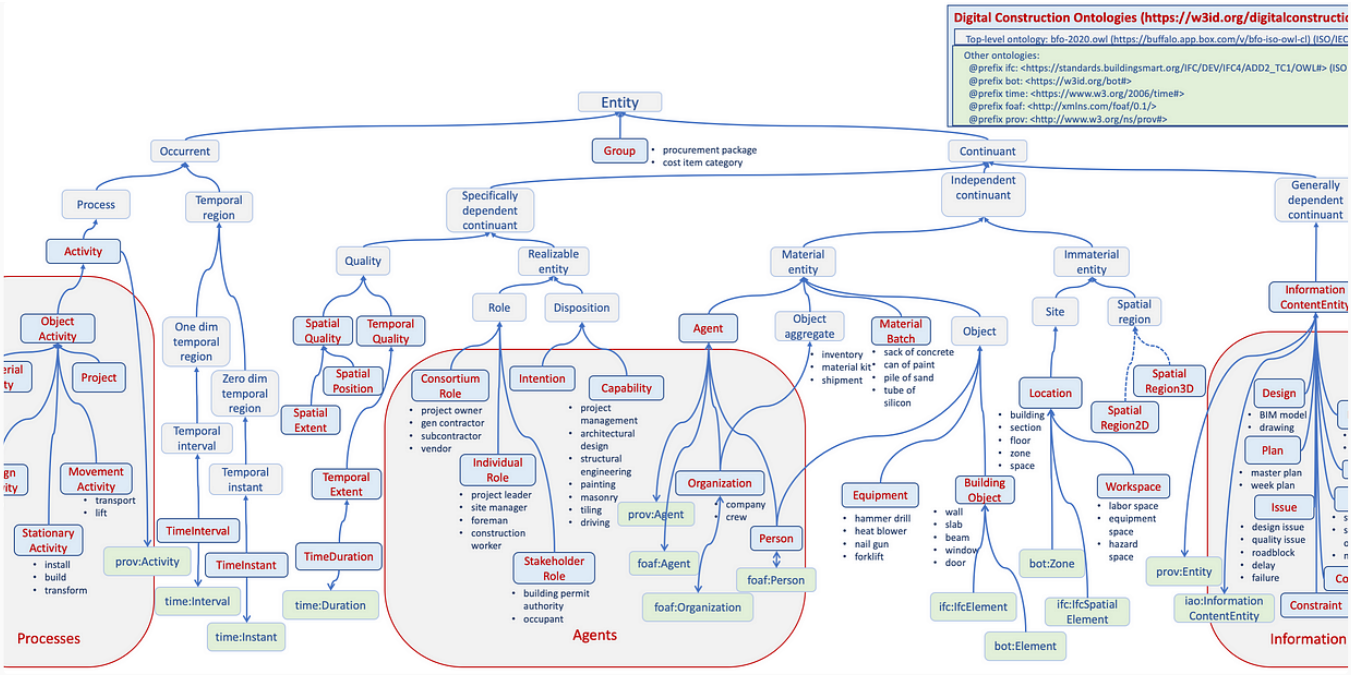
Alireza Dehbozorgi

Unraveling the Mystery of Prolog: Why it's More Relevant than Ever

Prolog, the Programming in Logic language, is a powerful tool for solving complex problems using a declarative approach. Prolog is a logic…

3 min read · May 19

4 1



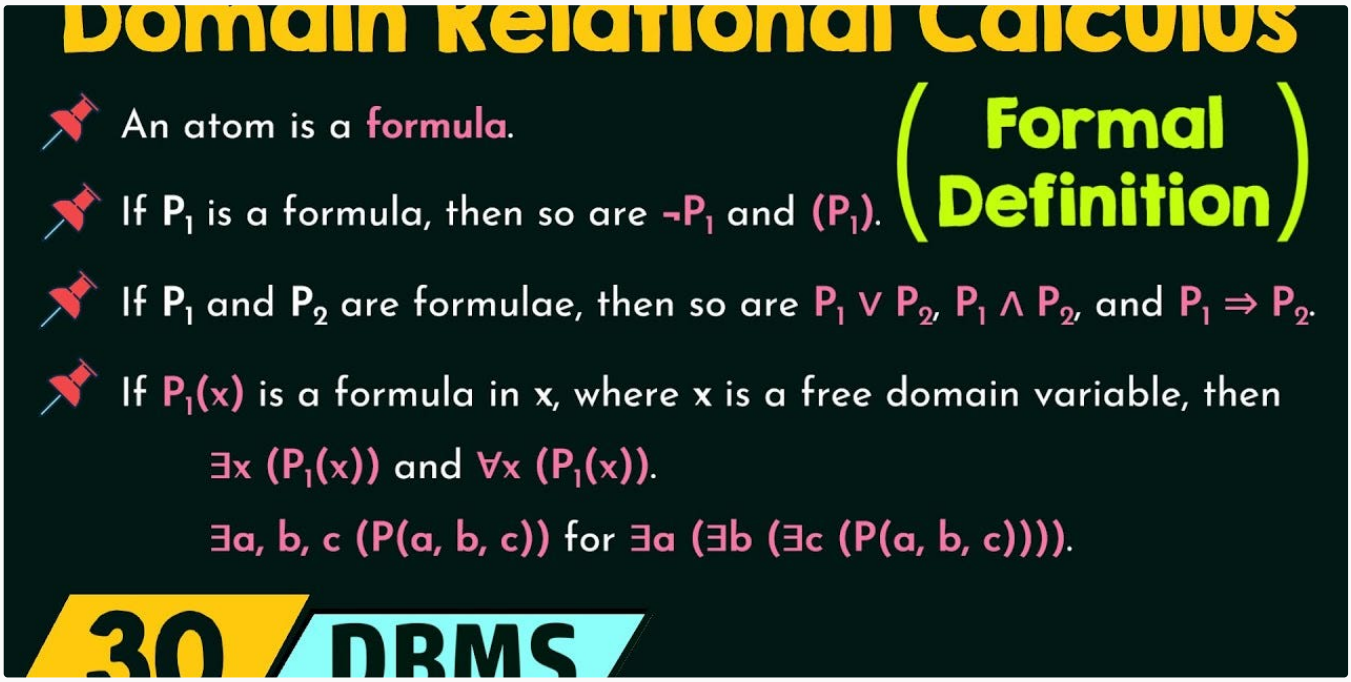
Alireza Dehbozorgi

Some Thoughts on Integrating Acoustic Speech Data into Formal Ontologies and Knowledge Management…

Acoustic analysis can be used to integrate speech cues and data into formal ontologies (1) by identifying and extracting acoustic features…

6 min read · 3 days ago

101 2



Alireza Dehbozorgi

Relational Calculus & Algebra (RCA) and Actionable Knowledge (Part I: Theoretical and Ethical…

Introduction

11 min read · May 11

1

Alireza Dehbozorgi

Computational Theories of Cognition- Part II: Logic-Based Models of Cognition.

This article explains the approach to reaching the overarching scientific goal of capturing the cognition of persons in computational formal…

18 min read · May 11

2

See all from Alireza Dehbozorgi

Recommended from Medium

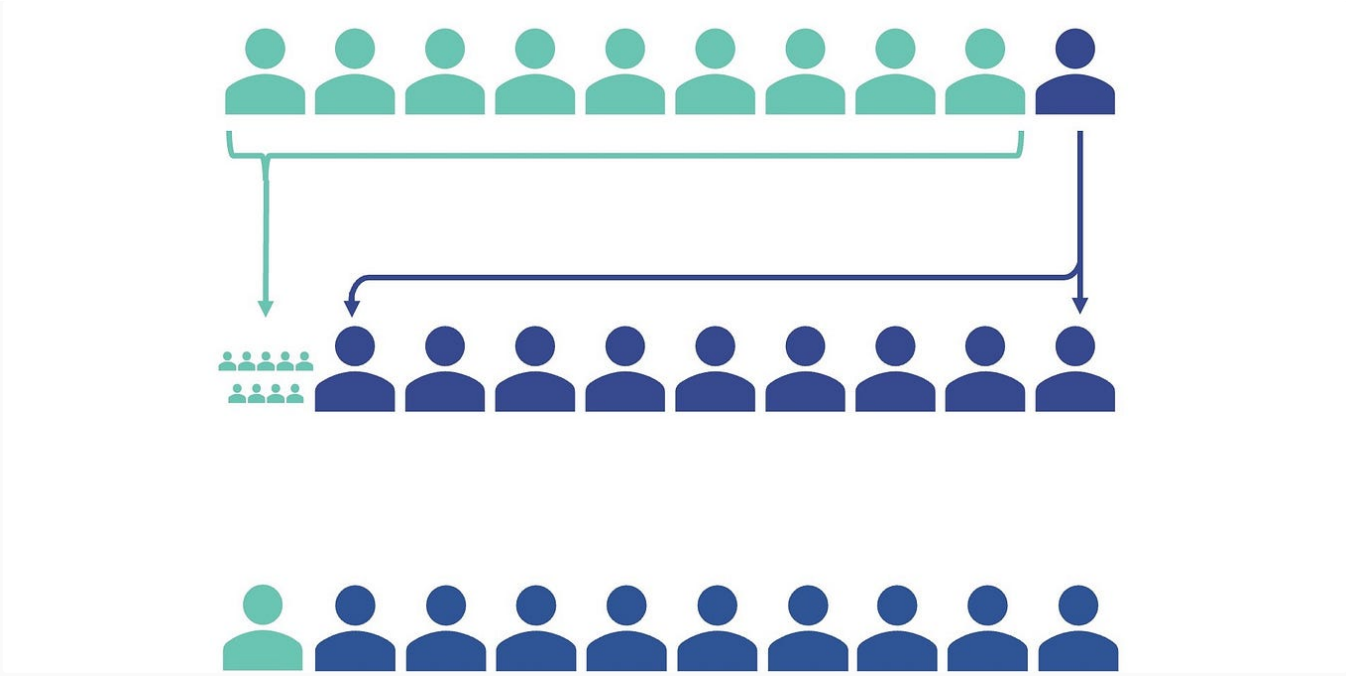
Chris Newman

A Love Letter to Gen Z

From your Big Brother, an Older Millennial

· 7 min read · Mar 10

4K 65



Matt Crooks in Towards Data Science

A Review of Propensity Score Modelling Approaches

A review of different approaches to using propensity scores in causal inference modelling

· 11 min read · May 17

108 1

Lists

What is ChatGPT?
9 stories · 52 saves

Stories to Help You Level-Up at Work
19 stories · 44 saves

Staff Picks
323 stories · 81 saves

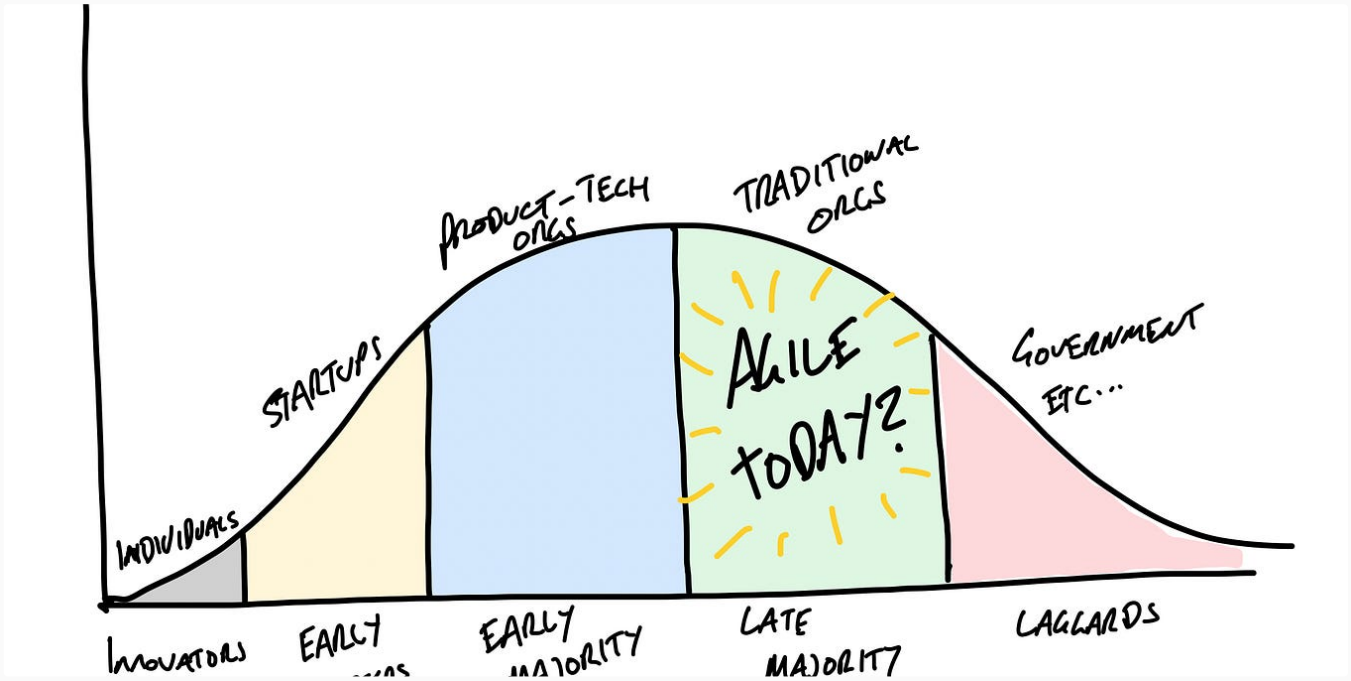
Rachel Greenberg in Entrepreneur's Handbook

Debunking 5 Pieces of Horrible Startup Advice

Just because a best-selling startup playbook tells you to do something doesn't make it right, relevant, or the optimal option.

· 11 min read · 2 days ago

698 14



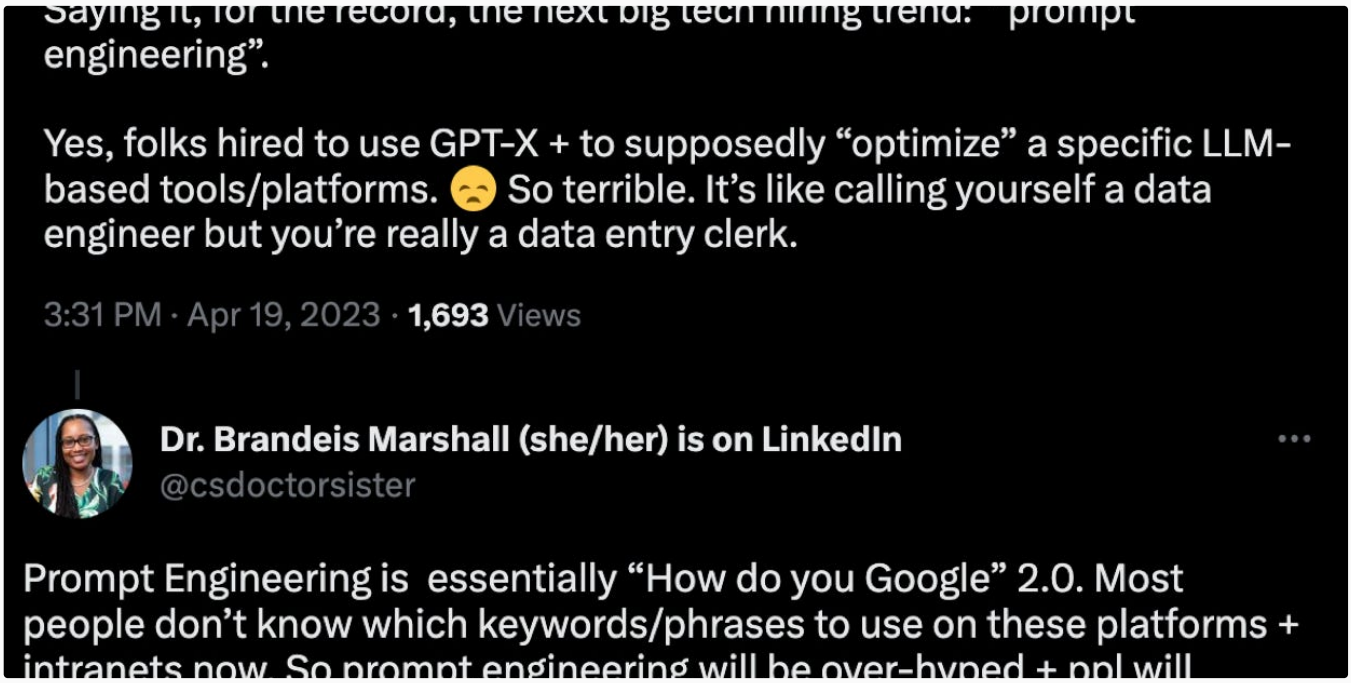
Anthony Murphy in Product Coalition

The Rise of Product-led Transformations

In 2019 an HBR article stated that of the $1.3 trillion spent on transformations, $900 billion was wasted.

· 8 min read · 4 days ago

258 11



Brandeis Marshall

What's UnAI-able

3 human-driven decision-making competencies that every industry needs

· 4 min read · 4 days ago

402 22

Neel Dozome in UX Collective

Times New Roman: can we make serifs great again?

Matthew Carter and the re-invention of type design for screens and desktop printing

· 30 min read · 21 hours ago

311 4