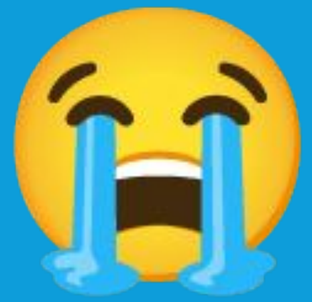
Diversifying Your Intent Training Data Library

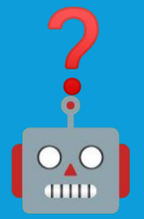
Chatbot Conference
April 12, 2022

Marlinda Galapon
Conversation Design Senior

salesforce design

The problem...



Biased training data → biased bots 

The solution...

Linguistically inclusive training data! →

Linguistically inclusive bots!

Inclusive Training Data

Building the industry’s first-ever inclusive training data set for chat-based conversational apps.

- **Prioritizing historically underrepresented varieties**
- **Building a library**
- **Setting a standard**
- **Drafting a framework**

Inclusive Training Data

Building the industry's first-ever inclusive training data set for chat-based conversational apps.

- **Prioritizing historically underrepresented varieties**
  - African-American English (AAE)
  - Chicano English
  - Southern U.S. English
  - English as a Second Language
- **Building a library**
- **Setting a standard**
- **Drafting a framework**

Image credit: Adobe Stock Photos

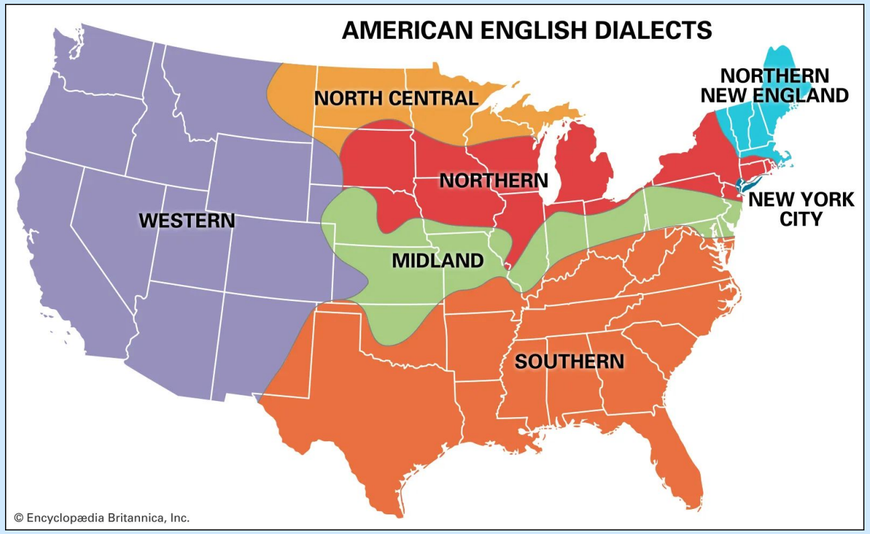


Image credit: Encyclopedia Britannica, Inc.

Inclusive Training Data

Building the industry's first-ever inclusive training data set for chat-based conversational apps.

- **Prioritizing historically underrepresented varieties**
  - African-American English (AAE)
  - Chicano English
  - Southern U.S. English
  - English as a Second Language

  Collected demographic data
  - gender identity, race, education level, age (which all affect language use)
- **Building a library**
- **Setting a standard**
- **Drafting a framework**

Image credit: Adobe Stock Photos

DESCRIPTIVE STATISTICS for Gender Identity				
Dialect	**Female**	**Male**	**Non-binary**	**Totals**
AAE	35	45	4	**84**
Chicano Engl.	9	6	0	**15**
ESL	32	9	1	**42**
U.S. Southern	51	31	4	**86**
Totals	**127**	**91**	**9**	**227**

Inclusive Training Data

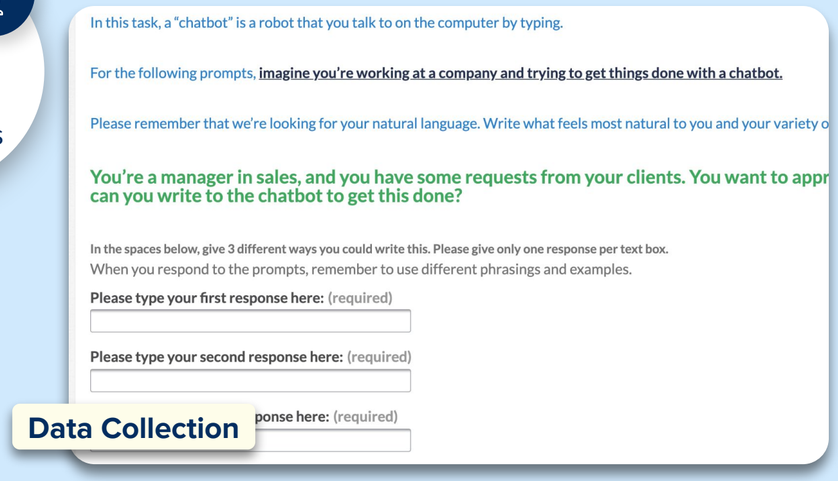
Building the industry's first-ever inclusive training data set for chat-based conversational apps.

- **Prioritizing historically underrepresented varieties**
- **Building a library of diverse training data that any team at Salesforce can easily draw from**
  - Research design
  - Data collection
- **Setting a standard**
- **Drafting a framework**

10,000+
Utterances collected

Image credit: Adobe Stock Photos

230
Research participants



Inclusive Training Data

Building the industry’s first-ever inclusive training data set for chat-based conversational apps.

- **Prioritizing historically underrepresented varieties**
- **Building a library of diverse training data that any team at Salesforce can easily draw from**
- **Setting a standard**
  - practical methods of ethical AI in practice
- **Drafting a framework**

Image credit: Adobe Stock Photos

Image credit: Adobe Stock Photos

Inclusive Training Data

Building the industry's first-ever inclusive training data set for chat-based conversational apps.

- **Prioritizing historically underrepresented varieties**
- **Building a library of diverse training data that any team at Salesforce can easily draw from**
- **Setting a standard**
- **Drafting a framework**
  - leveraging sociolinguistics to responsibly collect diverse data
  - Intentional selection of linguistic varieties
  - Informed by standardized sociolinguistic field methods

Image credit: Adobe Stock Photos

Image credit: Adobe Stock Photos

Inclusive Training Data

Building the industry’s first-ever inclusive training data set for chat-based conversational apps.

Impact:

Truly meets our users where they are—making our bots flexible enough to understand their dialect without issue, increasing our overall total addressable market across all areas of the company.

Ethics by Design

salesforce design

Thank You.

mgalapon@salesforce.com