



Diversifying Your Intent Training Data Library

Chatbot Conference
April 12, 2022

Marlinda Galapon
Conversation Design Senior

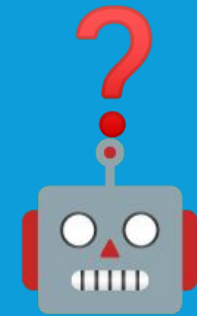




The problem...



Biased training data → biased bots





The solution...



Linguistically inclusive training data! →

Linguistically inclusive bots!



YAY!

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**



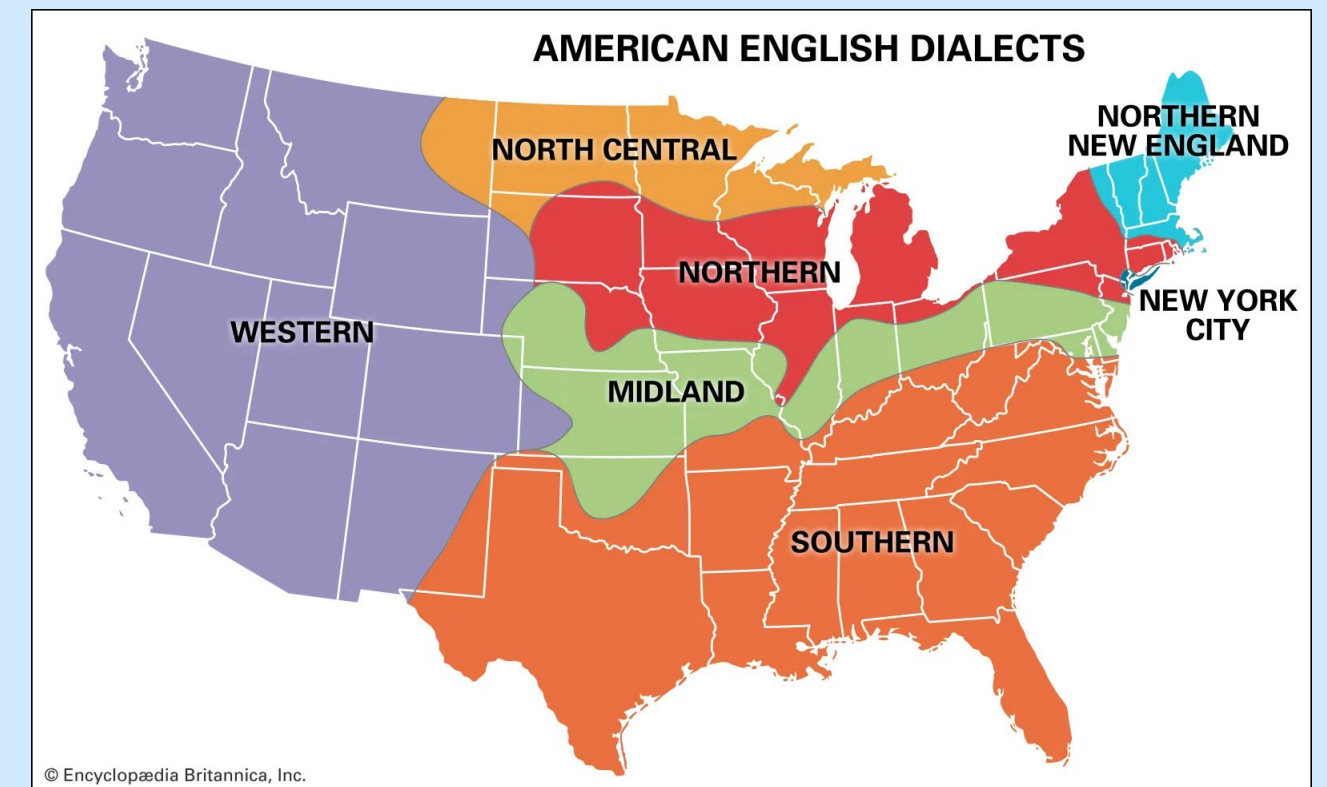
Inclusive Training Data

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



© Encyclopædia Britannica, Inc.

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

230
Research
participants



Image credit: Adobe Stock Photos

In this task, a "chatbot" is a robot that you talk to on the computer by typing.

For the following prompts, imagine you're working at a company and trying to get things done with a chatbot.

Please remember that we're looking for your natural language. Write what feels most natural to you and your variety of

You're a manager in sales, and you have some requests from your clients. You want to approve a request. How can you write to the chatbot to get this done?

In the spaces below, give 3 different ways you could write this. Please give only one response per text box. When you respond to the prompts, remember to use different phrasings and examples.

Please type your first response here: (required)

Please type your second response here: (required)

Please type your third response here: (required)

Data Collection

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



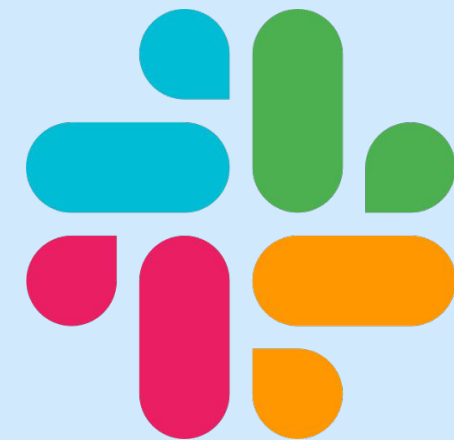
Image credit: Adobe Stock Photos

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



Thank You.



mgalapon@salesforce.com

