Language-tagging for contact language varieties: A Texas German case study

Margaret Blevins (mblevins@utexas.edu)
Department of Germanic Studies
The University of Texas at Austin

THE PROBLEM

1. Transcribed tokens of spoken contact language varieties have a lot of (orthographic) variation.
   → It can be difficult to find what you're looking for
   → Existing natural language processing tools (e.g., part-of-speech (POS) taggers) will often be less accurate because they are usually trained using data from a single (standard) language.
2. Token-level annotations such as orthographic normalization and POS-tags rely on interpretations with respect to language. What should these interpretations be based on and how can they be made transparent?

For example, the same transcribed token could be orthographically normalized differently, depending on what language it is presumed to be:

Transciption: *mir* ***ham*** ***ham*** *gegessen*
[ˈham] (German) /ˈhæm/ (English)

Orthographic normalization: wir **haben** **ham** gegessen
'we ate ham' / 'wir haben Schinken gegessen'

GOALS

- Develop token-level language-tagging system
- Make rationlle behind orthographic normalization and POS decisions transparent & reproducable
- Allow researchers to search for foreign material on the token level

METHOD & CORPUS

DATA

According to Boas (2009:34) **Texas German** is "a set of varieties of German spoken in Texas which have descended from the dialects of German brought to Texas in the 19th century."

This case study is based a set of excerpts from open-ended conversations in the Texas German Dialect Archive (TGDA). It encompases ~13 hours of conversation, and is proportionally representative for the first 600 speakers interviewed by the Texas German Dialect Project, with respect to birth location and gender.

TOOLS

The data was annotated using the EXMARaLDA Partitur-Editor (Schmidt 2016).

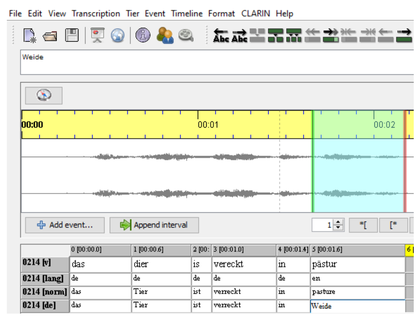


Fig. 1. Screenshot of the EXMARaLDA Partitur-Editor

DESIDERATA FOR THE ANNOTATION SYSTEM

- Clear, understandable guidelines that are freely accessessable → transparent & reproducable
- Relatively flexible & silmultaneously consistent
- Compatable across different languages and research paradigms

WHAT 'LANG' IS *NOT*

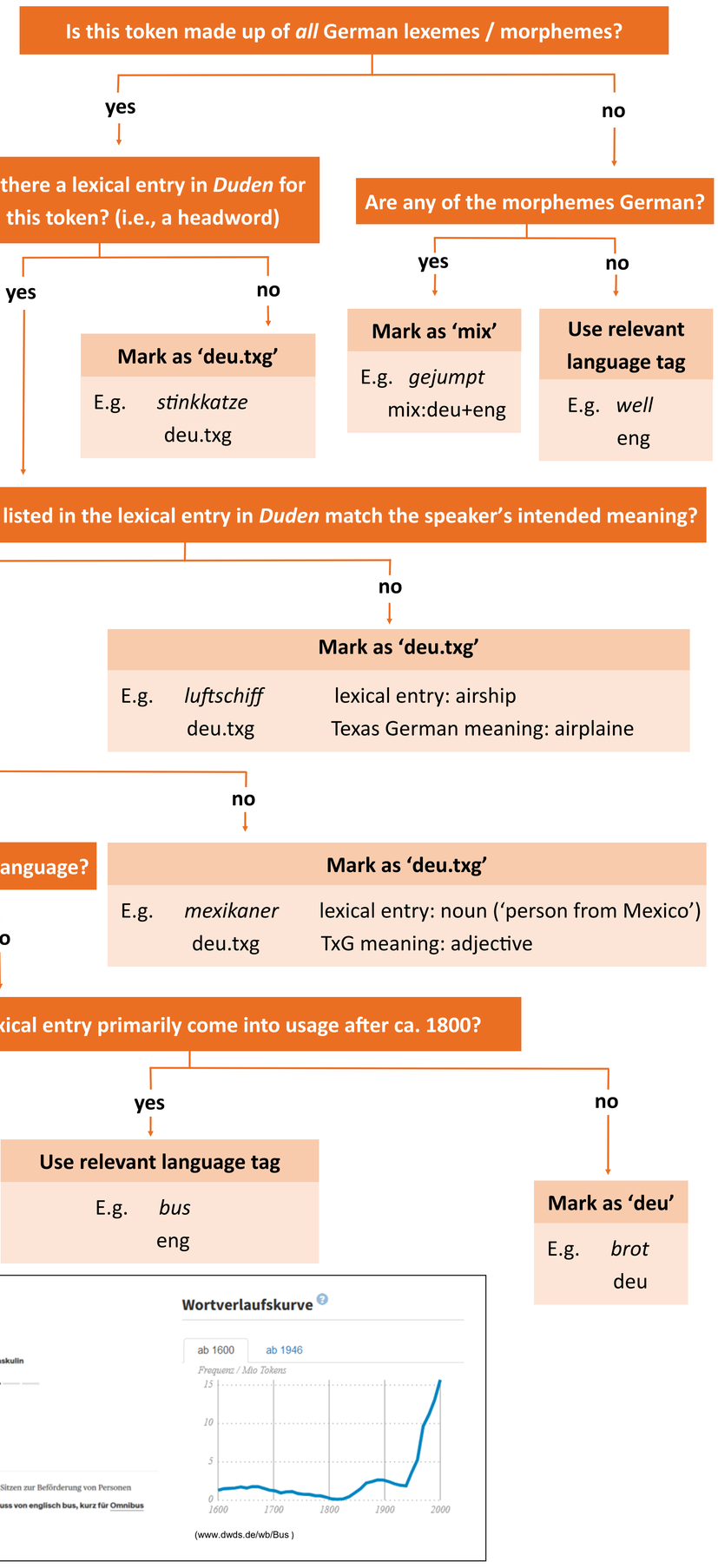
- The original etymological source of a lexeme, or else ...
  – English *cotton* would be tagged as Arabic and orthographically normalized as *quṭun*
  – German *Fenster* 'window' as would be tagged as Latin and normalized as *fenestra*
- Speaker-specific – a given token, if it represents the same semantic and morphological meaning, is language-tagged in the same way no matter who produces it or what idiolectal tendencies a speaker may have
- 'What language would this speaker consider this word to be?' – i.e., does the speaker consider the lexeme in question to be a German or an English word
- A categorization of loan word vs. code-switch vs. borrowing, etc.

LANGUAGE TAGGING SYSTEM

PRIMARY 'LANG' TAGS

Tag	Meaning
deu	German
eng	English
spa, wen, ces	Spanish, Wendish, Czech
mix:LANG+LANG	
deu.txg	
*	ambiguous

Language tags follow the **ISO 639-2** guidelines and decisions are primarily based on:
- morphological / lexical choice
- part-of-speech
- semantics



DISAMBIGUATING COGNATES

There are several ways of approaching cognates. All three of the following approaches are used in this system, depending on the situation.

Create a system to disambiguate cogntes

If a feature can be used to make a distinction between cognates, e.g., phonology, it should be used. For example, German *Musik* and English *music* phonologically differ from ech other in several respects, which makes differentiating between the two easier.

tok	musik	music
IPA	[muˈziːk]	/ˈmjuzɪk/
lang	deu	eng

Always make the same decision

There are certain tokens that are always ambiguous in the same way and are therefore always language-tagged and normalized in the same way, e.g., *denn* 'then' (standard German *dann*).

tok	und	denn		
lang	deu	deu		
NOT	deu	*deu	*deu.txg	*eng

If the phonology of two items is too similar, it may be appropriate to always mark it as ambiguous.

Example 1: Non-standard use of plural *–s*

tok	zwei	esels	
lang	deu	*deu	*mix:deu+eng

Example 2: Semantic shift of German lexeme vs. phonological adaptation of English lexeme (*Grad* meaning 'grade (in school)')

tok	grad	
lang	*deu.txg	*eng

Base the decision on context

There are certain tokens that have a great deal of overlap with regards to orthographical form, meaning, and pronunciation in English and German, e.g., the preposition *in.* In this case, instead of always tagging *in* as German or English, or always marking it as ambiguous, the language token is reliant on the language of the tokens to the immediate left and right of the in token. If the in token is surrounded by German tokens, it is marked as German. If it is surrounded by English tokens, it is tagged as English. If a German token is on one side, and an English token on the other, it is tagged as language ambiguous.

REFERENCES

Blevins, Margaret (2022): The language-tagging and orthographic normalization of German-language contact varieties. Dissertation, University of Texas, Austin.

Boas, Hans C. 2009. *The life and death of Texas German*. Durham: Duke University Press.

Boas, Hans C., Marc Pierce, Karen Roesch, Guido Halder, & Hunter Weilbacher. 2010. The Texas German Dialect Archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics* 22 (3), 277-296.

Schmidt, Thomas. 2016. EXMARaLDA Partitur-Editor. Manual. <https://www.exmaralda.org/pdf/Partitur-Editor_Manual.pdf>