The Denglisch Corpus of German-English Code-Switching

Doreen Osmelak

Language Science and Technology
Saarland University

Shuly Wintner

Computer Science
University of Haifa

6th May, 2023 SIGTYP @EACL 2023, Dubrovnik

Introduction

Code-Switching (CS):

Process of mixing two or more languages within a discourse or even within a single utterance

A User vor 4 Std.

Dass die Königin der Niederlande eine Lead Figure in diesem Konflikt war.

33 Antworten Teilen •••

Introduction

Code-Switching (CS):

Process of mixing two or more languages within a discourse or even within a single utterance

A Reddit vor 1 Std.

Ich will eine Foodtour durch ganz Berlin machen.

Antworten Teilen •••

Introduction

Code-Switching (CS):

Process of mixing two or more languages within a discourse or even within a single utterance

A Reddit vor 7 Std.

Ah, the Staatsangehörigkeitserwerbsurkunde.
You may now legally eat Bratwurst.

87 Antworten Teilen •••

Introduction

Code-Switching (CS):

Process of mixing two or more languages within a discourse or even within a single utterance

A User vor 3 Std.

Und trotzdem wird auf uns Europäern rumgebasht.

Damned if you do, damned if you don't

Antworten Teilen •••

Introduction

Code-Switching (CS):

Process of mixing two or more languages within a discourse or even within a single utterance

Phenomenon:

- still relatively little understood
- still few resources

Research so far:

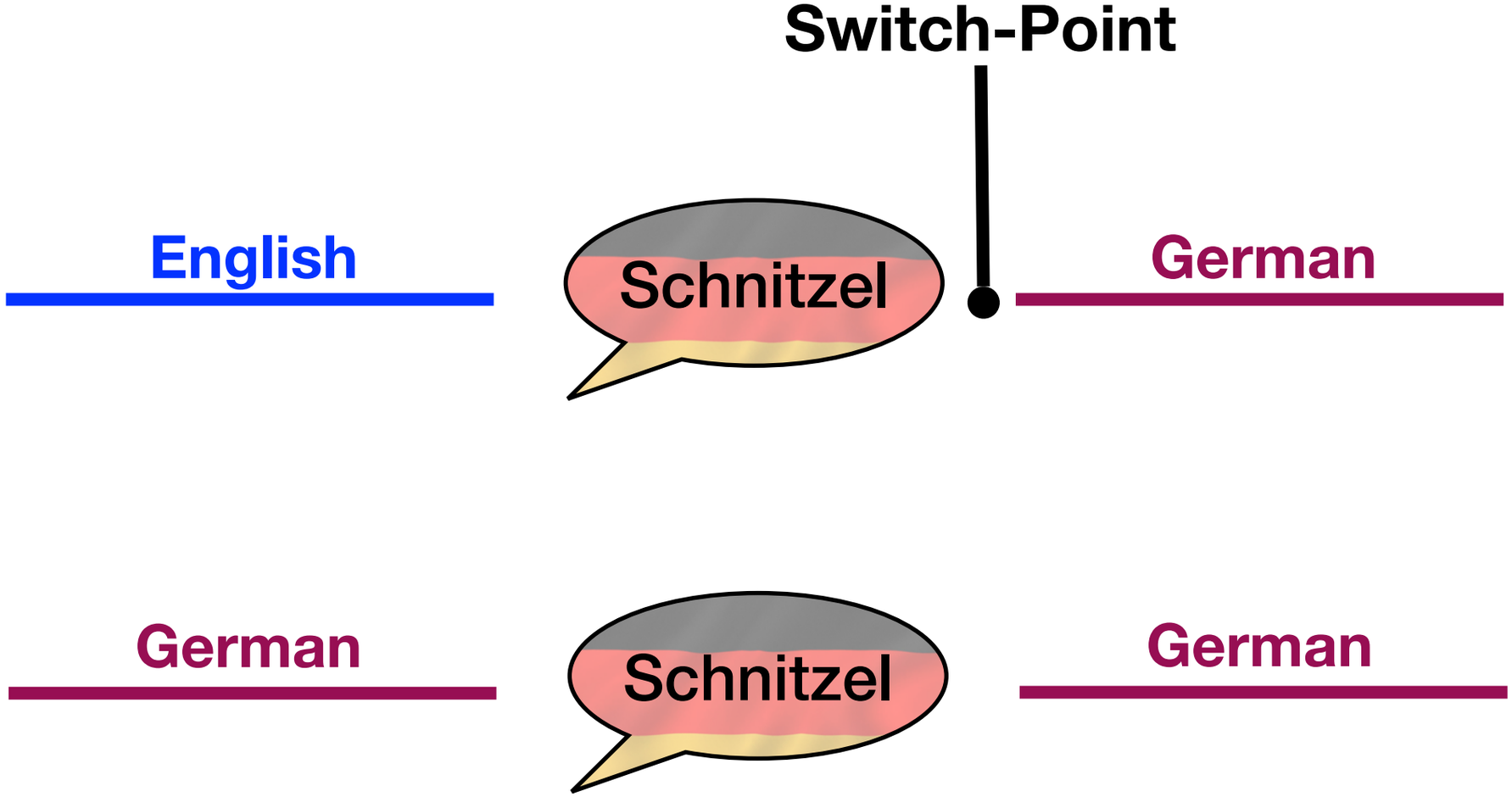
- focused on oral CS
- very limited data, small number of authors

Motivation

Triggering Hypothesis:

"Lexical items that can be identified as being part of more than one language for the speaker [...] may facilitate a transversion from one language to another" (Clyne 2003)

(→ Lexical transfer, bilingual homophones, proper names)



Overview

- annotation scheme for German-English
  incorporating shared words and origin

- corpus of German-English social media posts

- classifiers that predicts language-ID

Annotation Scheme

1	English				
2	German				
3	Overlaps				
	3a Named Entities		3c Merge-Words		3b Ambiguous Words
		3a-E English Origin		3c-C Compounds	
		3a-D German Origin		3c-M Morphology	3-E Untranslatable English
		3a-AE Adapted to English		3c-EC Entity Compounds	3-D Untranslatable German
		3a-AD Adapted to German		3c-EM Entity Morphology	3-O Untranslatable Other
4	Neutral				
	4a Foreign	4b Numbers		4d Interjections	<url> URL
		4b-E English only		4d-E English only	<punct> Punctuation
		4b-D German only		4d-D German only	<EOS> End of Sentence
		4c Smiley		4e-E English abbr.	<EOP> End of Paragraph

Overlaps: belong to both mental lexicons

Neutral: Language-universal
/ belong to all mental lexicons

special markers for origin of shared words (E/D/O)

Annotation Scheme

Named Entities:

Unadapted entities: → *3a-E / 3a-D*

Paris, Berlin

Translated entities: → *1 / 2*

United Kingdom -- Vereinigtes Königreich

Germany -- Deutschland

Orthographic adaptions: → *same as original entity*

*Constan**ce** -- **K**onstan**z***

Morphologic adaptions: → *3a-AD / 3a-AE*

Kalifornien -- California

Lexical adaptions: → *3a-AD / 3a-AE*

***New** Zealand -- **Neu**seeland*

Annotation Scheme

Borrowings:

Established untranslatables: → *3-E/D/O*

schnitzel, cheeseburger, döner

Unestablished untranslatables: → *1 / 2*

Blockchain Lockdown

Translatables: → *1 / 2*

Display -- Bildschirm

Integrated Old Loans: → *1 / 2*

cemetery, assassin, cotton

Unintegrated Old Loans: → *4a / 1*

PS e.g.

Neologisms and pseudo-borrowings: → *3-E/O*

video Handy (cellphone)

Annotation Scheme

Mixes:

Compounding:

→ *3a-E / 3a-D*

*Wohlstands**bubble** (prosperity bubble)*

Flexion

→ *3a-E / 3a-D*

ver**bugged, **rumge**bash**t

Entity Compounds:

→ *3a-E / 3a-D*

*NRA-**mäßig** (NRA-like)*

Inflected Entities:

→ *3a-E / 3a-D*

*googl**en** (to google)*

Language on Neutral Items:

cues to language: *90s--90er* → *4b-E/D*
ähm--erm → *4d-E/D*

Injection-like abbreviations: *lol, rofl* → *4e-E*

Annotation Scheme

Very few instances on some classes

E	English	1, 4b-E, 4d-E
D	German	2, 4b-D, 4d-D
M	Mix	3c, 3c-C, 3c-M, 3c-EC, 3c-EM
SE	Shared English	3a-E, 3a-AE, 3-E, 4e-E
SD	Shared German	3a-D, 3a-AD, 3-D
SO	Shared Other	3, 3a, 3b, 3-O, 4a, 4d
O	Other	4, 4b, 4c, <punct>, <url>

Corpus Creation

Corpus	Sentences	Strict CS	Relaxed CS	Posts with CS
Manually-tagged	4,200	1,250	1,400	950
Automatically-tagged	228,800	72,250	74,000	30,150
Total	233,000	73,500	75,400	31,100

Method from Rabinovich et al. (2019)

Manual annotation:

German-language subreddits (e.g. r/Berlin, r/DE)

Automatic annotation:

diverse range of German subreddits

Identifying Switches

- CRF sequence to sequence classifier

Features:

Orthography

Character N-grams

(Flexion/Derivation) Morphemes

Function Words

Frequency Lists

Lexical Components

Word Lists

Identifying Switches

Overall accuracy: 0.965

Tag	Prc	Rcl	F1	Support
English	0.97	0.98	0.98	29918
German	0.96	0.98	0.97	29730
Mix	0.50	0.19	0.28	246
Shared English	0.82	0.55	0.66	699
Shared German	0.78	0.54	0.64	807
Shared Other	0.75	0.50	0.60	1108
Other	0.99	0.98	0.99	12505
Micro Avg	0.96	0.96	0.96	75013
Macro Avg	0.82	0.68	0.74	75013
Weighted Avg	0.96	0.96	0.96	75013

Conclusion

- corpus of German-English CS utterances
  from user generated social media content

- precise language annotation, indicating switches

- addressed challenges in multilingual data by introducing
  various types of shared and mix categories

Future Work

- psycholinguistic research:
  investigate correlation btw Shared Items and Switches

  → forthcoming

Thank You