Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT FÜR
GEISTESWISSENSCHAFTEN

INTRODUCTION

Why "Code-Switching" in Corpus Linguistics?

- Perpetual occurrence of CS in the multilingual online community
- Structural pattern as groundwork for analysing social motivation
- Distinction of different types of CS
- Creating training data: CS is a great challenge for many natural language processing applications such as machine translation, speech recognition, and information extraction.

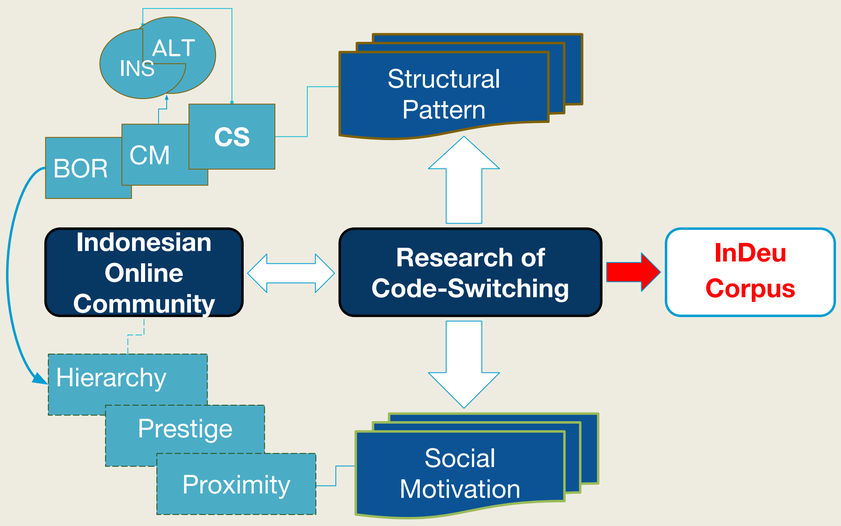


Fig. 1: The structural and sociolinguistic perspective in the CS research

Why Online Community in "kaskus.co.id"?

- Feature language is mixed: Indonesian and English
- Unique English internet jargons and terms in Kaskus dictionary
- High tendency of language mixing in forums
- Focus of the study: Kaskus forum category “Regional” (subforum “Europe”, subsubforum “Germany”)
- “Germany” is the largest active Kaskus community in Europe
- Status ranking of Kaskus users in a hierarchy

The Terms of Mixed Language

Borrowing (BOR):
The adaptation of a lexical element from the embedded language without its grammatical aspects, and when it has a high usage frequency by the speakers of the dominant language (cf. Myers-Scotton 2006, 1993).

Code-Mixing (CM):
“All cases where lexical items and grammatical features from two languages appear in one sentence” (Muysken 2000, p.1). CM is sub-classified into alternation (ALT) and insertion (INS):

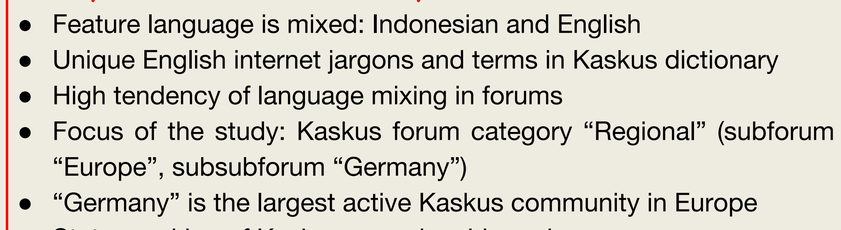


Fig. 2: The patterns of alternation and insertion (Muysken 2000, p. 7)

Code-Switching (CS):
“The juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems” (Gumperz 1982, p. 59), whereas the speaker uses CS for a certain purpose or pragmatic function respectively (cf. Androutsopoulos 2011).

Hypothesis, Research Questions and Aims

The annotation of CS provides evidence for its discourse functions (DF), particularly as a strategy to keep intimacy and to show prestige.

1. Which constructions of code-switching occur in the Kaskus posts?
2. Which discourse functions are conveyed by the code-switching in each topic of the thread?

Structural approach
To distinct evidently BOR, CM and CS (as well as INS from ALT) due to frequency and grammar/morphosyntax elements
Sociolinguistic approach
To derive the DF from the annotated CS elements (INS and ALT), based on its pragmatic function

THE InDeu CORPUS

- Manual download
  - 32 threads from kaskus.co.id (25 single-page; 7 multi-page)
  - Comprising 586 posts: ~60,000 tokens (~6,200 types)
- Manual annotation in EXMARaLDA (exmaralda.org)
- POS tagset: DEU (STTS, 1999), Lang3 (Penn Treebank, 1990), IND (for this poster: POS TAG - University of Indonesia, 2014)

	jack,	fressen	biasanya	dilakukan	oleh	binatang/hewan,
Lang	ENG	DEU				
POS	NE-ITJ	VVINF	RB	VB	IN	NN
Lang Mix Type	CS<INS: DF5					
Translat	Jack, „fressen“ is usually used for animal,					

	kalau	essen	itu	Manusia,
Lang		DEU		
POS	SC	VVINF	PR	NN
Lang Mix Type		CS<INS: DF4<ITER		
Translat	but „essen“ is for human,			

	masa	sih	Betreuer	ngomong	gitu	ke	elu?
Lang			DEU				
POS	RB	RP	NN	VB	PR	IN	NN
Lang Mix Type			CM				
Translat	really, the supervisor said that to you?						

Fig. 3: The main categories of annotation in the InDeu corpus on EXMARaLDA

- Creating CS tagset based on Myers-Scotton (2006, 1993), Muysken (2000), Androutsopoulos (2011):

BOR

1) Institutional or Scientific Terms in DEU: Uni; FH; studkol ‘Studienkolleg’; ABH ‘Ausländerbehörde’; Arbeitnehmer; Firma; etc.
2) Orthographic changing in Lang3: en (and), konek (connect), skul (school), nubi (newbie), trid/trit (thread), etc.

CM

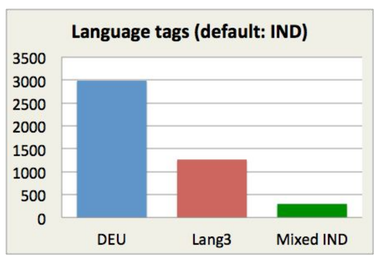
1) No equivalent in semantics or habituality: Wohnung; Heizung; Aufenthaltserlaubnis; Preisleistungsverhältnis; relax; perfekt; etc.
2) Morphosyntactic: *di***support** (supported); **penner**2 (Pennerinnen);
Kayaknya lu kurang gaul dengan orang yg **deutsch***nya bagus*.
ADV+ENKL PRP RB VB SC-IN NN KON NN+ENKL ADJA
[it seems you don’t hang out with the people, who speak good German]

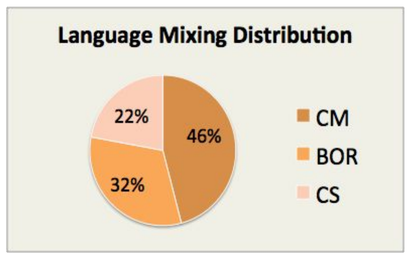
CS (the most used DF: “example”)

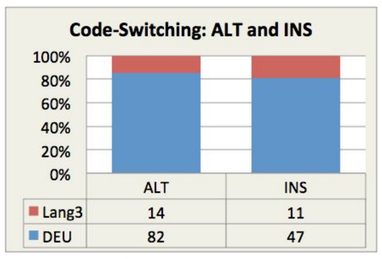
INS	ALT
Iteration: “Ausländer”; “Glück”. **Topic-comment**: “hochqualifizierter Ausländer” (topic); “Doktorarbeit” (comment). **Proximity**: formal routine “Hallo”, “danke”. **Distance**: “complain”; “you copy paste *aja* [easily] link”; “*pake fasilitas* [use the facility of] search”; Niederlassungserlaubnis.	**Formal routine**: “Danke im Voraus”; “Viel Erfolg”; “Herzlichen Glückwunsch”. **Direct speech**: “*..lgs berpikir* [directly think], OMG, i love this country..” (cf. fig. 8 in ‘DISCUSSION’). **Emphasizing through specific phrase**: idiom “der klügere gibt nach”, joke (fig. 3); **Proximity**: through joke (fig. 3)

Fig. 4: Analysis of social motivation from the structural patterns

Annotation statistics







Inter-annotator agreement
- Test file: 2,476 tokens
- All lang tags including default IND: Cohen’s kappa = 0.91
- 9 fine-grained lang (mixing) tags excluding IND (562 tokens): Cohen’s kappa = 0.59

Fig. 5: Analyses of manual annotation of languages and language mixing types

- Conversion to a column format for automatic processing

EXB-ID	TOKEN	POS	LANG	MIX	FILE	COMMENT
1793	10-09-2010	NA	NA	NA	3B.4b.3-11	meta: date
1806	jack	NE-ITJ	eng	CS<INS	3B.4b.3-11	
1806	,	$,	NA	NA	3B.4b.3-11	
1807	fressen	VVINF	deu	CS<INS	3B.4b.3-11	DF5
1808	biasanya	RB	ind	NA	3B.4b.3-11	
1809	dilakukan	VB	ind	NA	3B.4b.3-11	

Fig. 6: Sample from InDeu corpus in column format

ANALYSIS

- The language tagging between DEU and Lang3 can be ambiguous
- DEU: Institutional terms; Lang3: Chat language or internet jargon
- The looser the topic of a thread is, the more grammatical failures/spelling errors can appear in the embedded language. Ex. of failures: “Plastik tute” (compound); “ich bin studieren” (lemma)
- Similar result as other related works: Nouns were inserted most frequently
- The Kaskus rankings as the hierarchy among the Kaskus community do not influence the CS usage in Kaskus. The higher the tendency of a Kaskus user to switch the codes, the higher their position among their interlocutors:

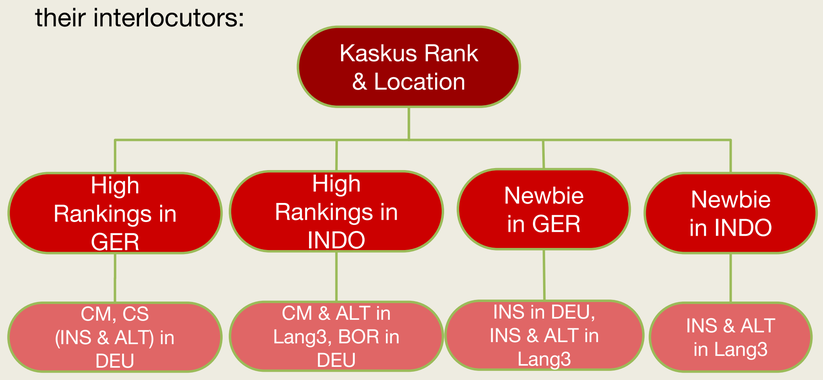


Fig. 7: Correlation between the status of Kaskus rank and the location of the Kaskus user

DISCUSSION

Can we solve these Problems?

Some problems during the annotation of the InDeu corpus by two annotators:

1) “Betreuer” (supervisor) is one of the characters in Hendy’s story, posted in the Kaskus-thread:

	Betreuer:	wie	so	Falsch?	Hendy:	Aber	gibt	es	[…]
Lang		DEU				DEU			
Translat	Supervisor: Why is it wrong?				Hendy: But, is there […]				

Fig. 8: The annotation of a character in the direct speech in a narrative

Comparing to “Betreuer” in fig. 3, does the “Betreuer” above has a function as a meta information? Or should we annotate the word in the language category as well as tagging its POS?

2) Language ambiguity between two annotators (ANO)
- 1st ANO: Bus/DEU-Lang3-IND; 2nd ANO: Bus/none
Solution: it is IND and not annotated.
- “Hi” in “**Hi**, schade*, apa kabar di* München” [Hi, Bummer, what’s up in München?]. 1st ANO: Hi/Lang3; 2nd ANO: Hi/DEU.
Solution: Hi/DEU-ENG. Do you agree with us about this?

REFERENCES

1. Androutsopoulos, J. 2011. Code-Switching in computer-mediated communication. In: Handbook of the Pragmatics of CMC.
2. Auer, P. ed., 2013. Code-switching in conversation: Language, interaction and identity. Routledge.
3. Çetinoglu, Ö. & Ç. Çöltekin. 2016. Part of Speech Annotation of a Turkish-German Code-Switching Corpus. In Proceedings of LAW-X.
4. Grosjean, F. 1982. Life with two Languages: an Introduction to Bilingualism. Harvard University Press.
5. Gumperz, J. 1982. Discourse strategies. Vol. 1. Cambridge University Press.
6. Indonesian POS tag - Tagger site (2014, October). Retrieved from URL: http://bahasa.cs.ui.ac.id/postag/tagger
7. Lemnitzer, L. & H. Zinsmeister. 2015. Korpuslinguistik, Eine Einführung. Tübingen: Narr Francke
8. Myers-Scotton, C. 1997. Duelling languages: Grammatical structure in codeswitching. Oxford University Press.
9. Myers-Scotton, Carol (2006): Multiple voices. An introduction to bilingualism. Malden, MA: Blackwell Pub.
10. Muysken, P. 2000. Bilingual speech. A typology of code-mixing. Cambridge University Press.
11. Vyas, Y. et al. 2014. POS Tagging of English-Hindi Code-Mixed Social Media Content. In Proceedings of EMNLP.

HANI PRIANDINI
E-Mail: priandini.hani@gmail.com
MARISKA AJENG HARINI
E-Mail: mariskaajeng@gmail.com
HEIKE ZINSMEISTER
E-mail: Heike.Zinsmeister@uni-hamburg.de