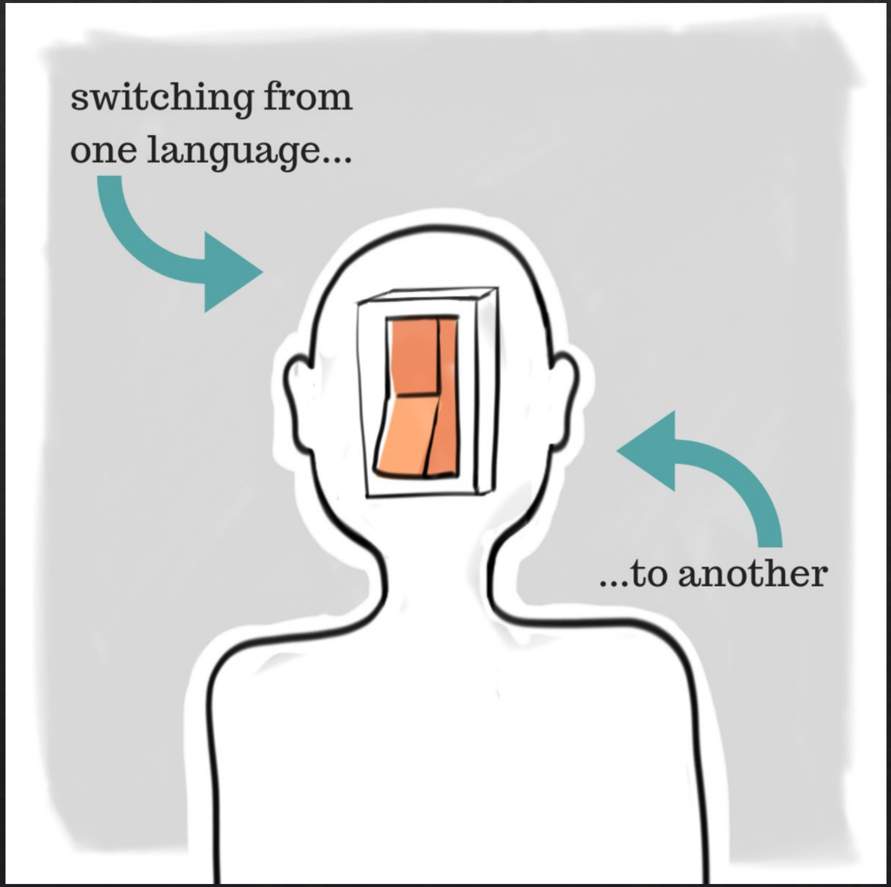
Code-Switching: Background, History, Data

601.764

3/9/2023

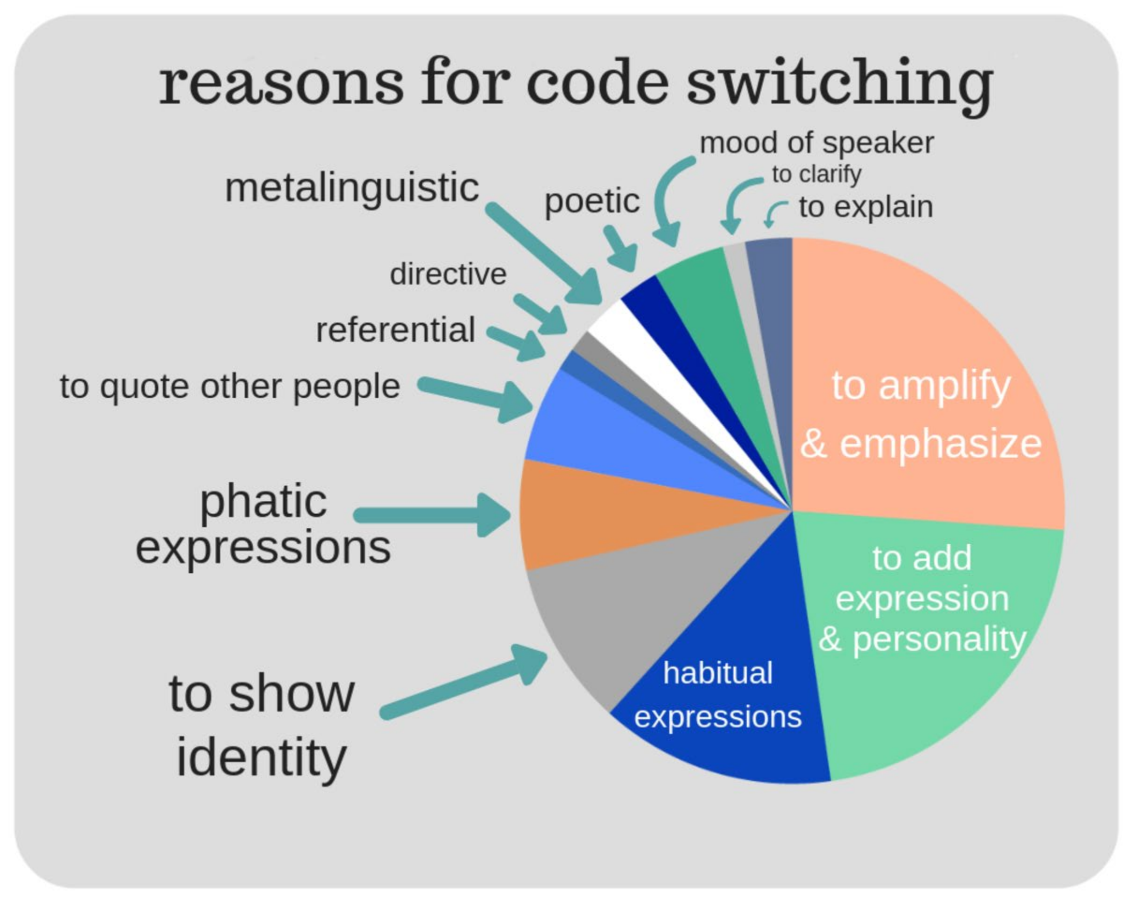
Definition?

- Code Switching
- Code-Switching
- Codeswitching
- Codemixing



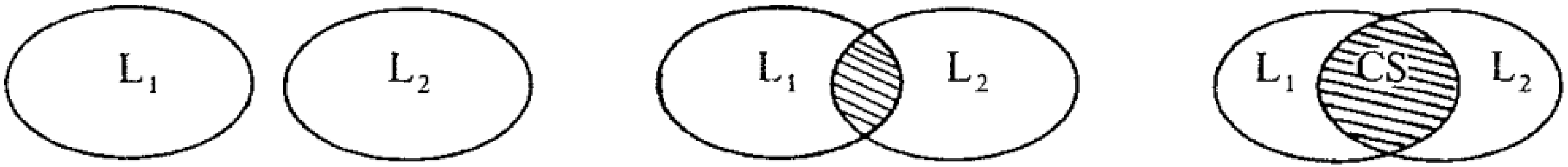
https://owlcation.com/humanities/Code-Switching-Definition-Types-and-Examples-of-Code-Switching

Grain of Salt:
No citations for this image....
Doubt ratios



https://owlcation.com/humanities/Code-Switching-Definition-Types-and-Examples-of-Code-Switching

Inter vs. Intrasentential

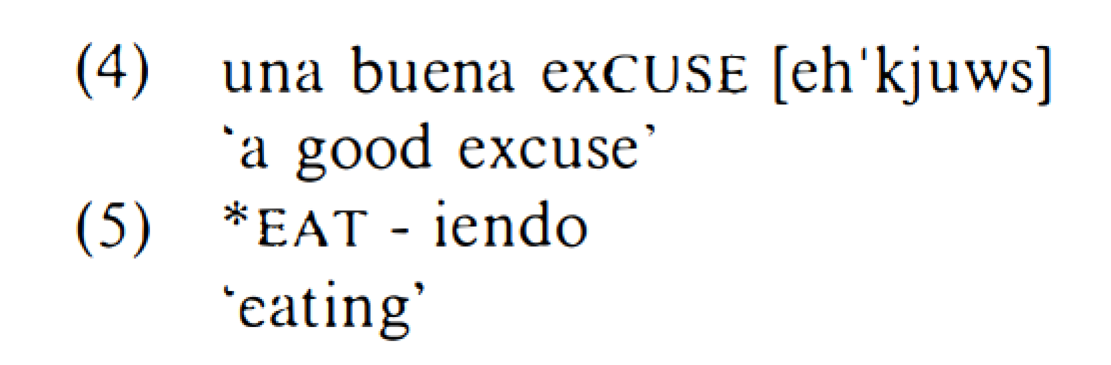


a. Inter-sentential switching b. ‘tag’-switching c. Intra-sentential switching

Insert a tag or short phrase:
He is famous, ya tú sabes.

Poplack 1980

What is allowed?



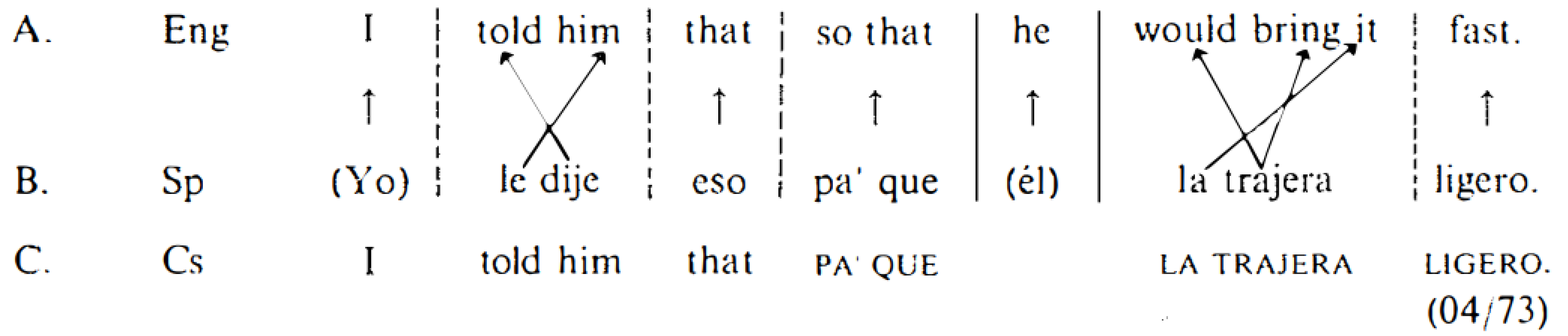
(4) una buena exCUSE [eh'kjuws]
‘a good excuse’

(5) *EAT - iendo
‘eating’

Do we think this is still true?

Poplack 1980

What is allowed?



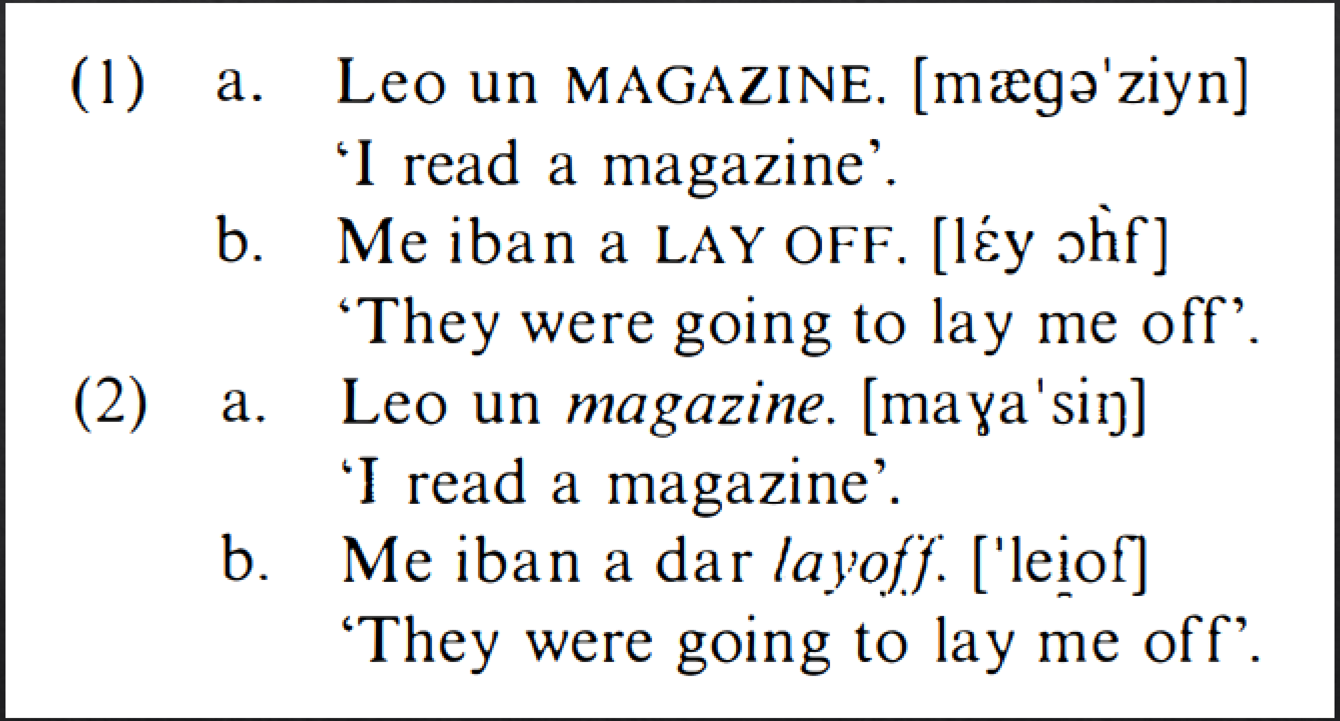
A.	Eng	I	told him	that	so that	he	would bring it	fast.
B.	Sp	(Yo)	le dije	eso	pa' que	(él)	la trajera	ligero.
C.	Cs	I	told him	that	PA' QUE		LA TRAJERA	LIGERO. (04/73)

Figure 1. *Permissible code-switching points*

Do we think this is still true?

Poplack 1980

Phonology as a boundary?



(1) a. Leo un MAGAZINE. [mægəˈziyn]
‘I read a magazine’.
b. Me iban a LAY OFF. [lɛ́y ɔh̀f]
‘They were going to lay me off’.

(2) a. Leo un *magazine*. [maɣaˈsiŋ]
‘I read a magazine’.
b. Me iban a dar *layoff*. [ˈlei̯of]
‘They were going to lay me off’.

Poplack 1980

History

- Most look at Blom and Gumperz 1972 as first study
- Norwegian Fishing Villiage

History

- Mexican Americans in Tucson, AZ
- Barker 1947 … early American study in linguistic anthropology
- "How does it happen, for example, that among bilinguals, the ancestral language will be used on one occasion and English on another, and that on certain occasions bilinguals will alternate, without apparent cause, from one language to another?"
- Family interactions → Spanish
- Formal interaction with Anglo-Americans → English
- …. Even if both were bilingual
- Less clearly defined situations → Less Fixed
- Younger people more like to use multiple languages
- Local Tucson identity
- Analysis from Nilep 2006

History

Amazon.com

Amazon.com: Languages in Contact: French, Ger
and Romansch in Twentieth-Century Switzerlan

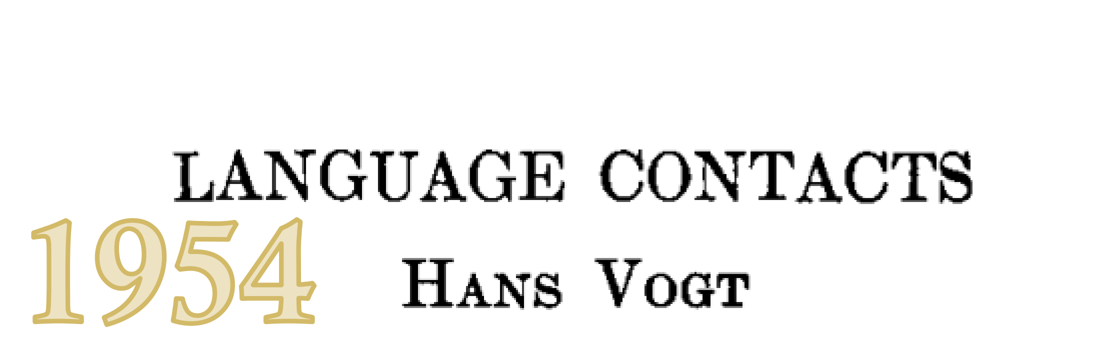
$149.00* · Out of stock

Languages in Contact
French, German and Romansh in twentieth-century Switzerland
Uriel Weinreich
With an introduction and notes by Ronald I. Kim and William Labov
John Benjamins Publishing Company
DE GRUYTER
Uriel Weinreich
LANGUAGES IN CONTACT
Deutsch
Französisch
Italienisch
Rätoromanisch
LANGUAGES IN CONTACT
Findings and Problems
URIEL WEINREICH
Mouton
URIEL WEINREICH
LANGUAGES IN CONTACT
Findings and Problems
401 W423l ej.2

By Tschubby - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=113480064

History

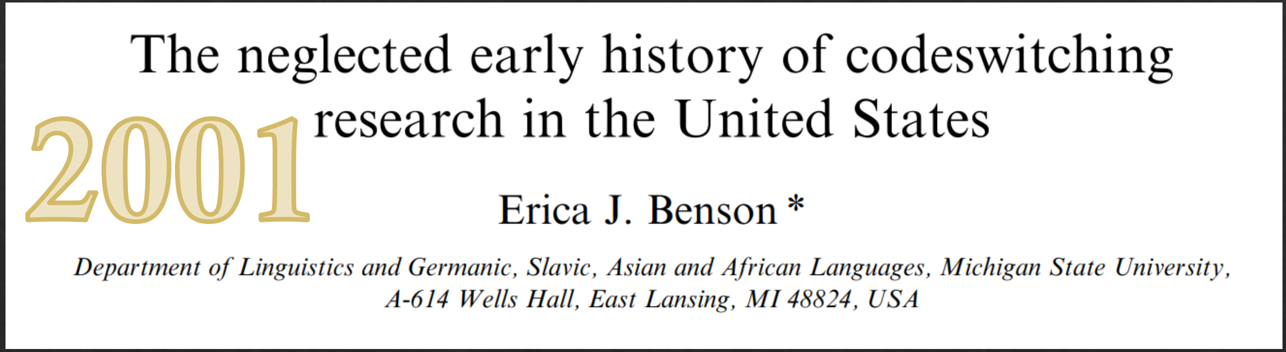
“Code-switching in itself is perhaps not a linguistic phenomenon, but rather a psychological one, and its causes are obviously extra-linguistic.”



LANGUAGE CONTACTS

1954

Hans Vogt



The neglected early history of codeswitching research in the United States

2001

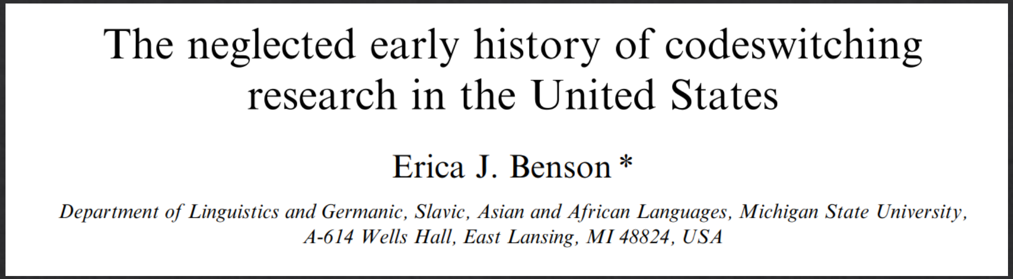
Erica J. Benson *

Department of Linguistics and Germanic, Slavic, Asian and African Languages, Michigan State University, A-614 Wells Hall, East Lansing, MI 48824, USA

“In linguistic theory new terminologies have been proposed in which terms such as systems and codes, patterns and structures, play a great part.”

“When languages are viewed as systems or codes, it becomes of primary interest to investigate what happens when linguistic systems come into contact with each other. The way bilingualism affects linguistic systems can be expected to throw light on the basic concepts we use in dealing with isolated systems”

“Ironically, it was in a review of Languages in contact (Vogt, 1954a) that I found the earliest (thus far) documented use of `codeswitching'. Vogt’s first instance of `codeswitching' was in reference to Meillet who he believed had little interest `in individual cases of codeswitching' (Vogt, 1954a)”



The neglected early history of codeswitching research in the United States

Erica J. Benson *

Department of Linguistics and Germanic, Slavic, Asian and African Languages, Michigan State University, A-614 Wells Hall, East Lansing, MI 48824, USA

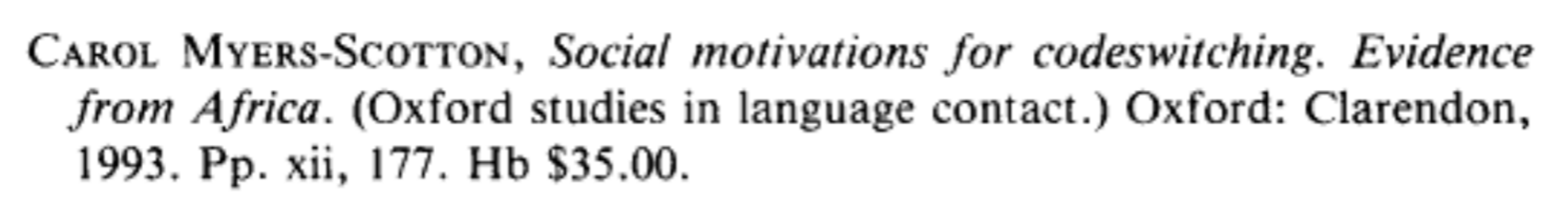
History

- “ [Weinreich] `the ideal bilingual switches from one language to the other according to appropriate changes in the speech situation (interlocutors, topics, etc), but not in an unchanged speech situation, and certainly not within a single sentence' and that some bilinguals have `a facility in switching languages even within a single sentence or phrase’.”
- “Weinreich labeled the phenomenon as `switching code' and referred the reader to Jakobson et al. (1952) and Fano (1950)”
- “Richard Diebold's (1962) presentation entitled `Code-switching in Greek-English bilingual speech'(which appears to be the ®rst publication to use `codeswitching' in the title).”

4 Classic Works

- The Norwegian language in America (1953) by Einar Haugen
- Bilingualism in the Americas (1956) by Einar Haugen
- Languages in contact (1953) by Uriel Weinreich
- Diglossia (1959) by Charles Ferguson.
- * According to Benson 2001

Matrix Language

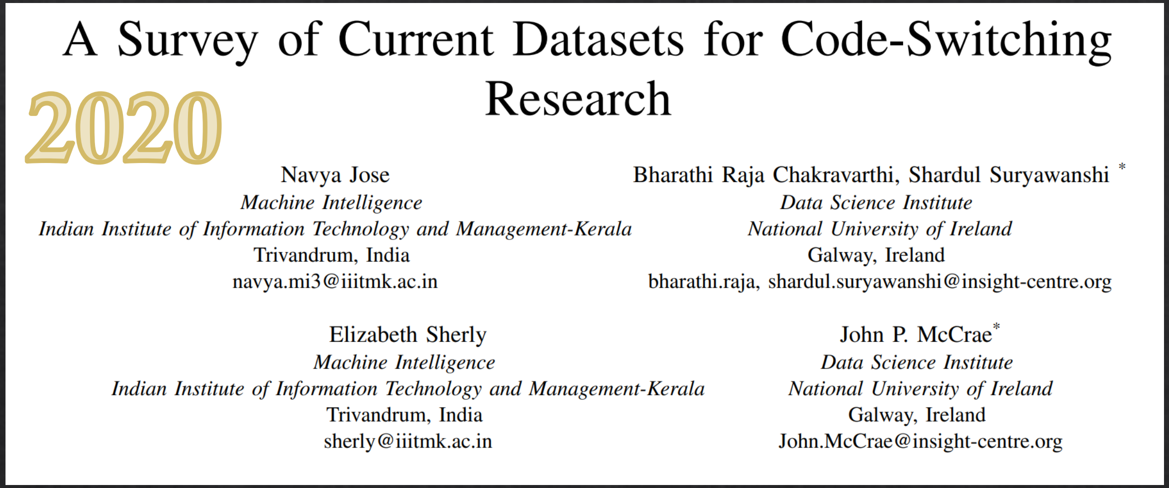


Carol Myers-Scotton, *Social motivations for codeswitching. Evidence from Africa*. (Oxford studies in language contact.) Oxford: Clarendon, 1993. Pp. xii, 177. Hb $35.00.

1993

Wakasa 2004

Datasets



A Survey of Current Datasets for Code-Switching Research

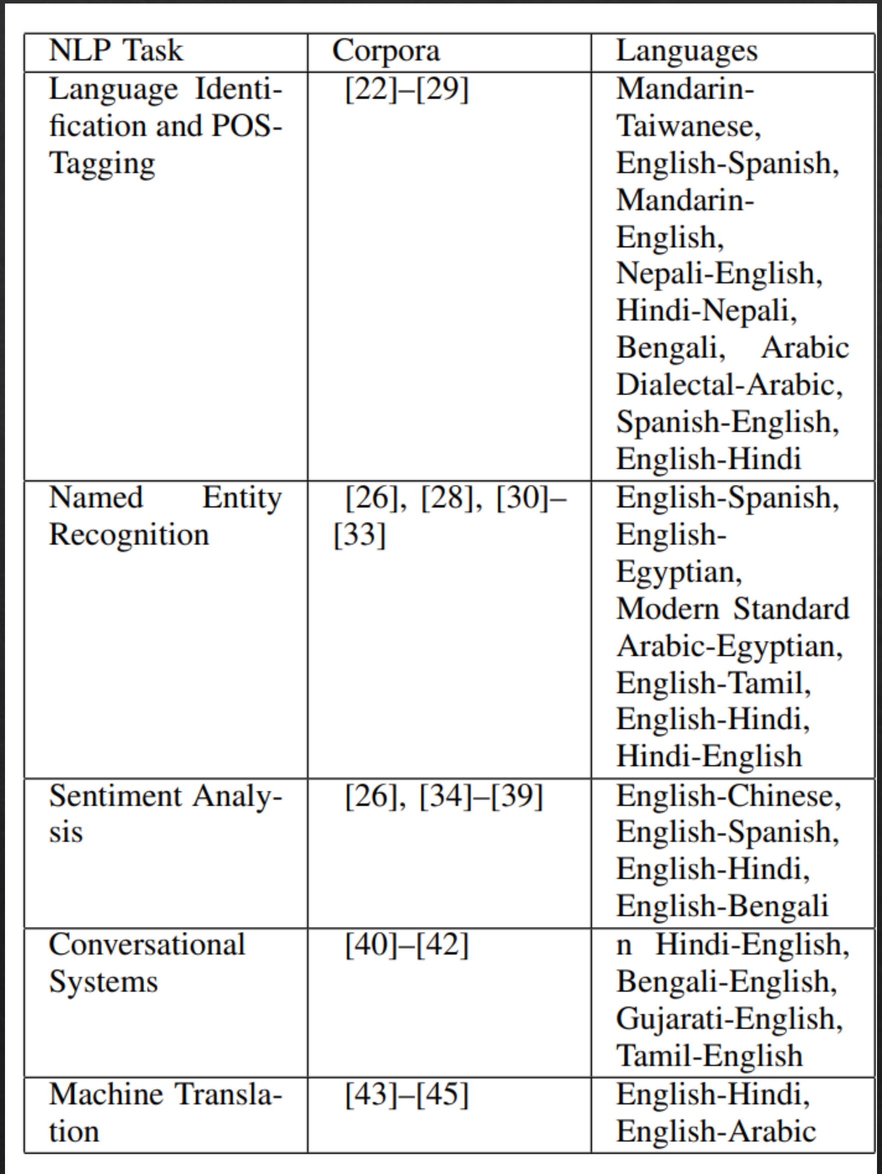
2020

Navya Jose
Machine Intelligence
Indian Institute of Information Technology and Management-Kerala
Trivandrum, India
navya.mi3@iiitmk.ac.in

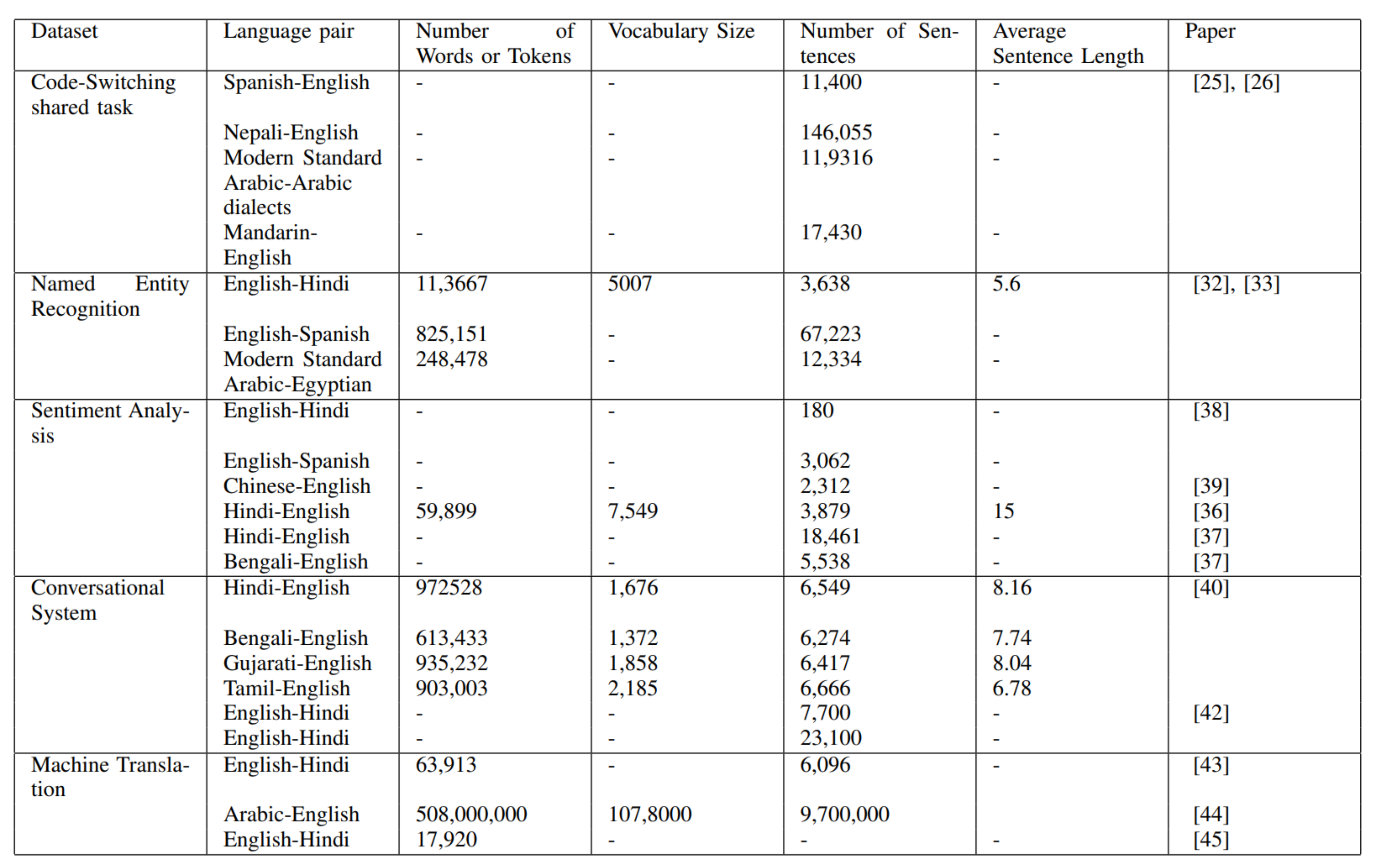
Bharathi Raja Chakravarthi, Shardul Suryawanshi *
Data Science Institute
National University of Ireland
Galway, Ireland
bharathi.raja, shardul.suryawanshi@insight-centre.org

Elizabeth Sherly
Machine Intelligence
Indian Institute of Information Technology and Management-Kerala
Trivandrum, India
sherly@iiitmk.ac.in

John P. McCrae*
Data Science Institute
National University of Ireland
Galway, Ireland
John.McCrae@insight-centre.org

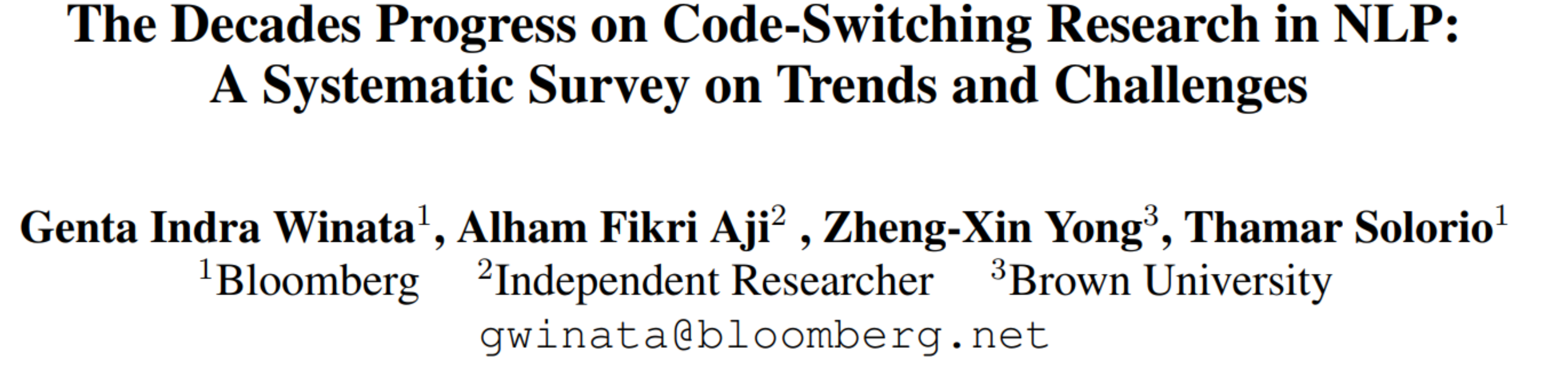


NLP Task	Corpora	Languages
Language Identification and POS-Tagging	[22]–[29]	Mandarin-Taiwanese, English-Spanish, Mandarin-English, Nepali-English, Hindi-Nepali, Bengali, Arabic Dialectal-Arabic, Spanish-English, English-Hindi
Named Entity Recognition	[26], [28], [30]–[33]	English-Spanish, English-Egyptian, Modern Standard Arabic-Egyptian, English-Tamil, English-Hindi, Hindi-English
Sentiment Analysis	[26], [34]–[39]	English-Chinese, English-Spanish, English-Hindi, English-Bengali
Conversational Systems	[40]–[42]	n Hindi-English, Bengali-English, Gujarati-English, Tamil-English
Machine Translation	[43]–[45]	English-Hindi, English-Arabic



Dataset	Language pair	Number of Words or Tokens	Vocabulary Size	Number of Sentences	Average Sentence Length	Paper
Code-Switching shared task	Spanish-English	-	-	11,400	-	[25], [26]
	Nepali-English	-	-	146,055	-	
	Modern Standard Arabic-Arabic dialects	-	-	11,9316	-	
	Mandarin-English	-	-	17,430	-	
Named Entity Recognition	English-Hindi	11,3667	5007	3,638	5.6	[32], [33]
	English-Spanish	825,151	-	67,223	-	
	Modern Standard Arabic-Egyptian	248,478	-	12,334	-	
Sentiment Analysis	English-Hindi	-	-	180	-	[38]
	English-Spanish	-	-	3,062	-	
	Chinese-English	-	-	2,312	-	[39]
	Hindi-English	59,899	7,549	3,879	15	[36]
	Hindi-English	-	-	18,461	-	[37]
	Bengali-English	-	-	5,538	-	[37]
Conversational System	Hindi-English	972528	1,676	6,549	8.16	[40]
	Bengali-English	613,433	1,372	6,274	7.74	
	Gujarati-English	935,232	1,858	6,417	8.04	
	Tamil-English	903,003	2,185	6,666	6.78	
	English-Hindi	-	-	7,700	-	[42]
	English-Hindi	-	-	23,100	-	
Machine Translation	English-Hindi	63,913	-	6,096	-	[43]
	Arabic-English	508,000,000	107,8000	9,700,000		[44]
	English-Hindi	17,920	-	-	-	[45]

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

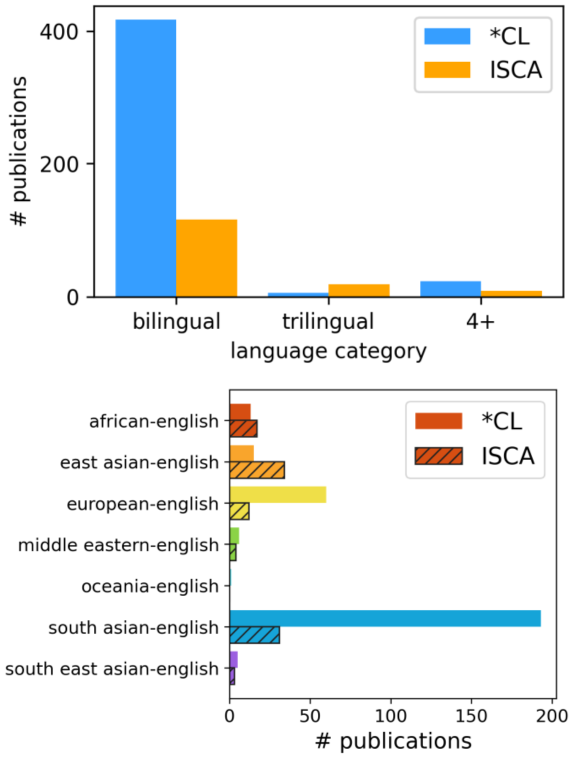


Figure 3: **(Top):** Number of publications across the type of language combination (bilingual, trilingual or 4+. **(Bottom):** Number of publications on fine-grained bilingual category with English as the L2 language.

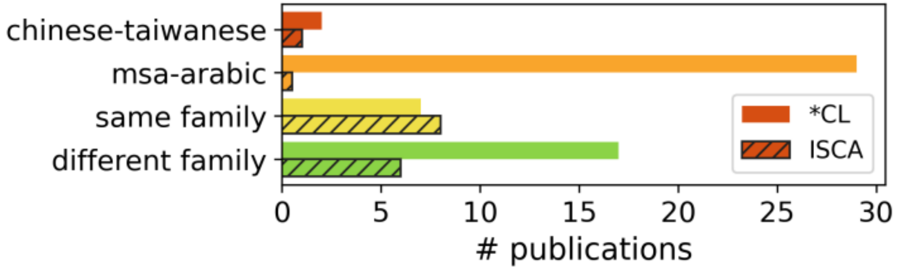


Figure 4: Number of publications of bilingual code-switched languages that do not contain English. *msa stands for Modern Standard Arabic. The first two are the combination of a language with its dialect.

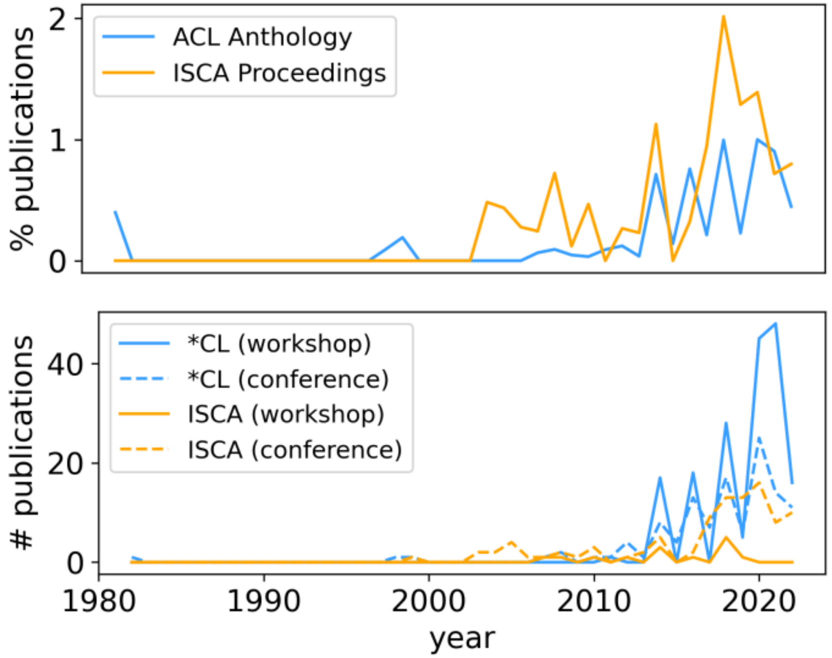
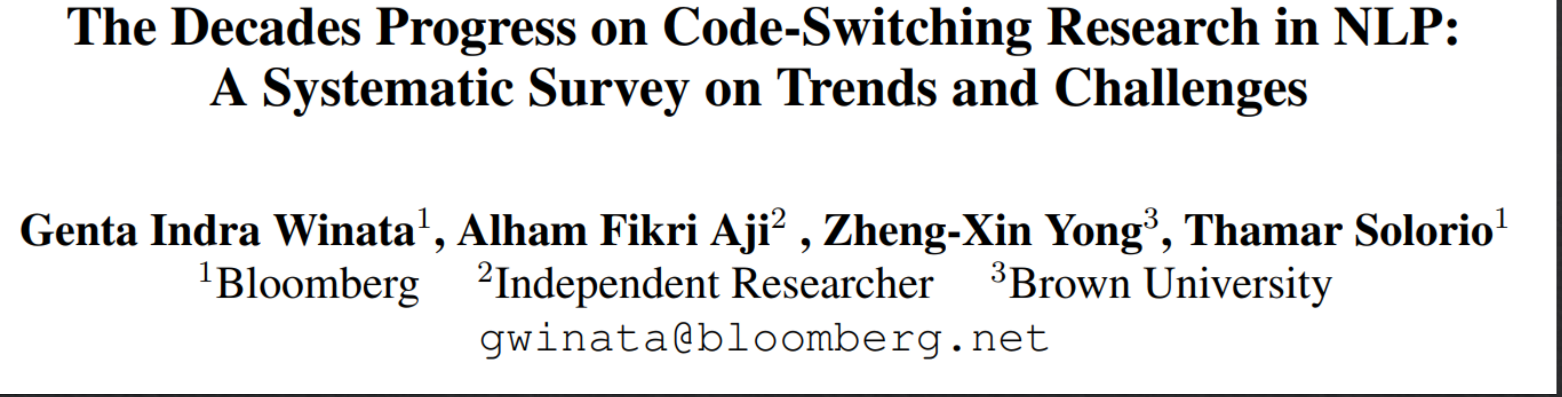


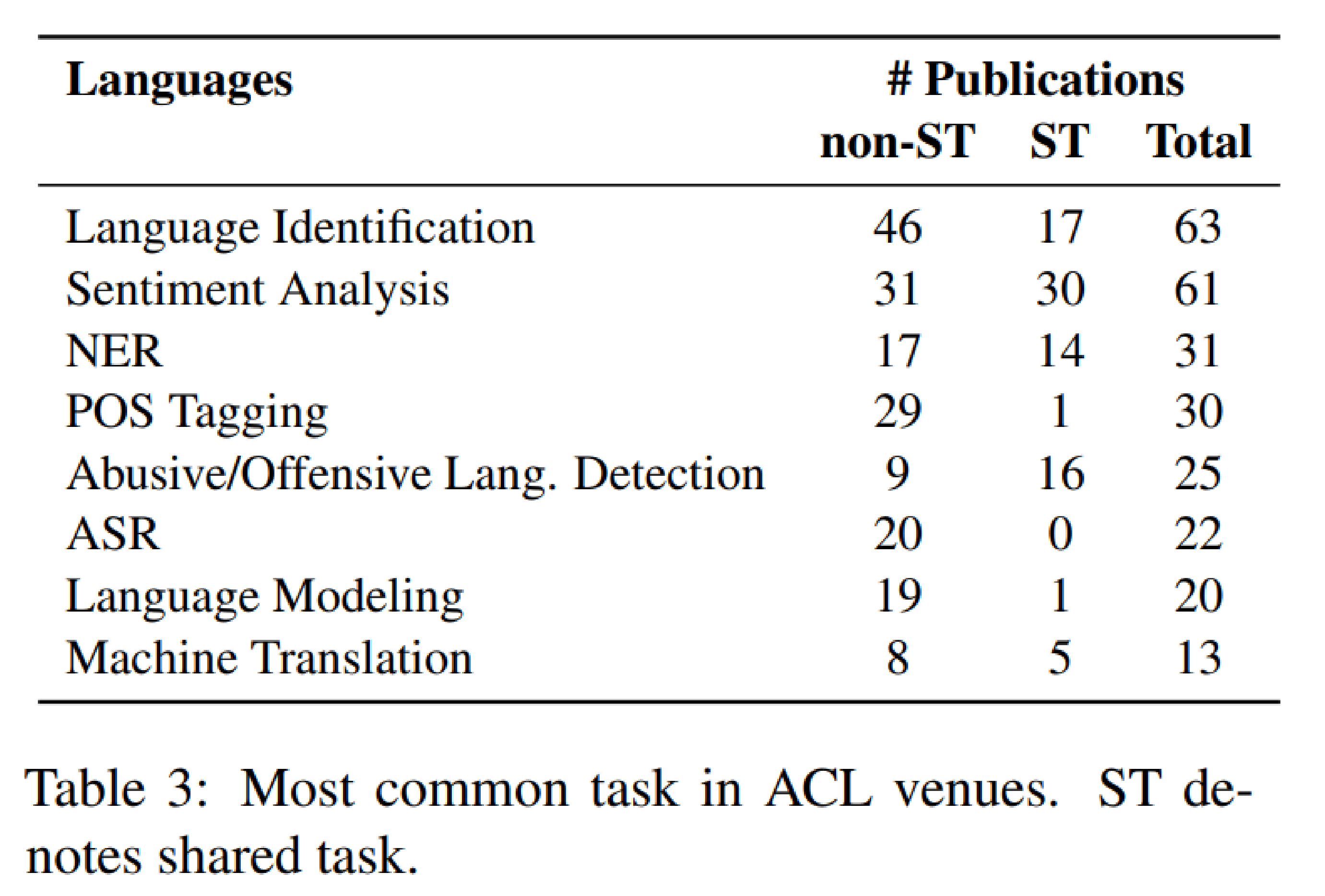
Figure 1: Number of publications over time in *CL and ISCA venues. We collect the papers on October 2022. **Top:** Relative to all *CL and ISCA papers. **Bottom:** absolute number, broken down into conferences vs workshops. It does not include papers published after. The graphs do not show the number of publications published in journals and symposiums.

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

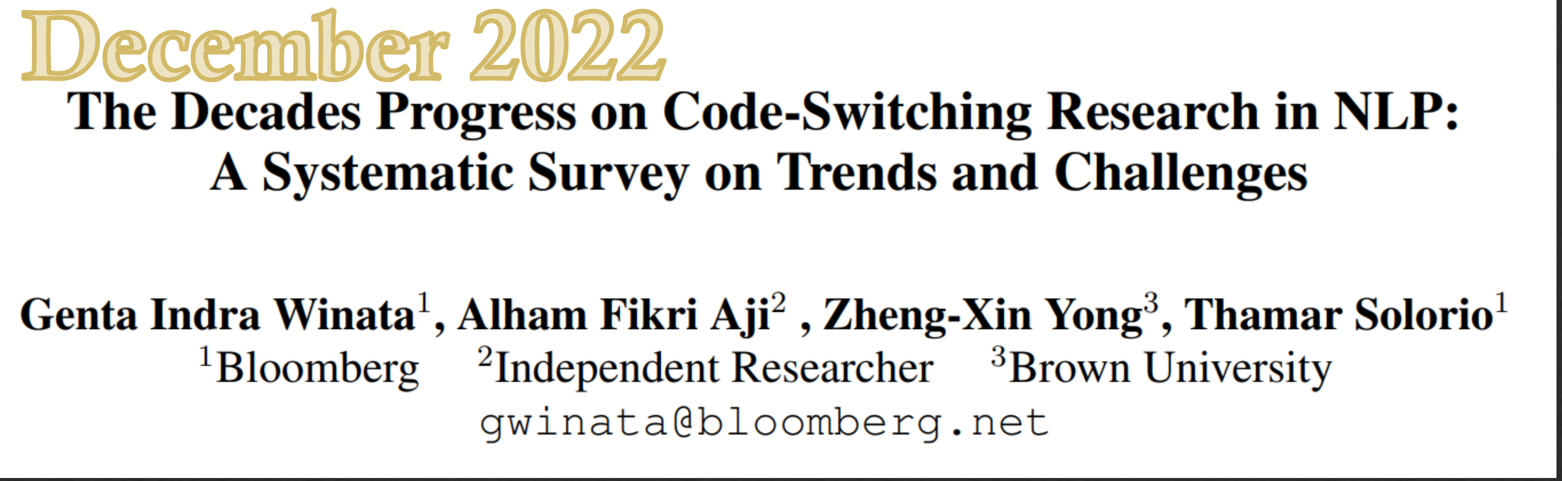
Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



Languages	# Publications		
	non-ST	ST	Total
Language Identification	46	17	63
Sentiment Analysis	31	30	61
NER	17	14	31
POS Tagging	29	1	30
Abusive/Offensive Lang. Detection	9	16	25
ASR	20	0	22
Language Modeling	19	1	20
Machine Translation	8	5	13

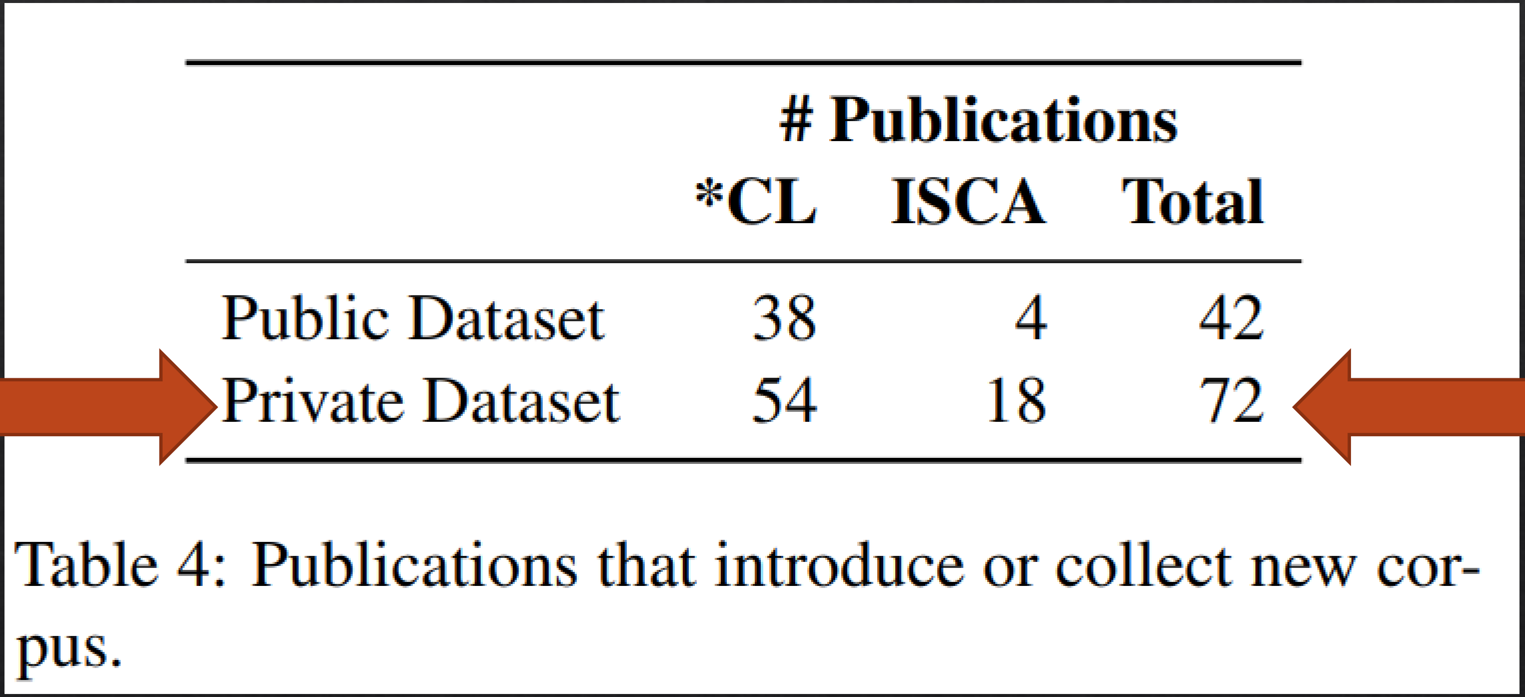
Table 3: Most common task in ACL venues. ST denotes shared task.

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

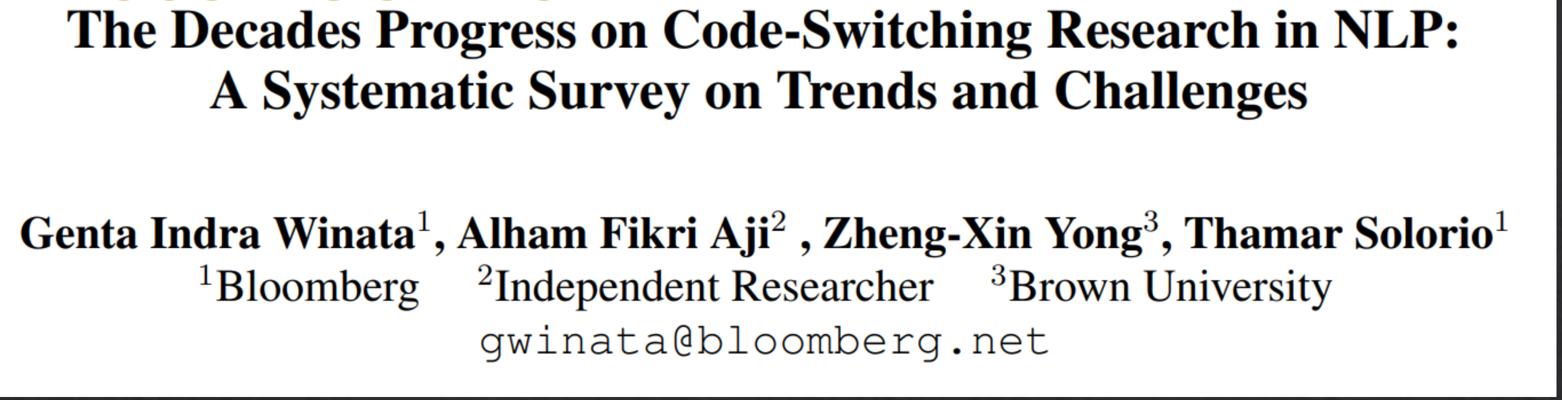


	# Publications		
	*CL	ISCA	Total
Public Dataset	38	4	42
Private Dataset	54	18	72

Table 4: Publications that introduce or collect new corpus.

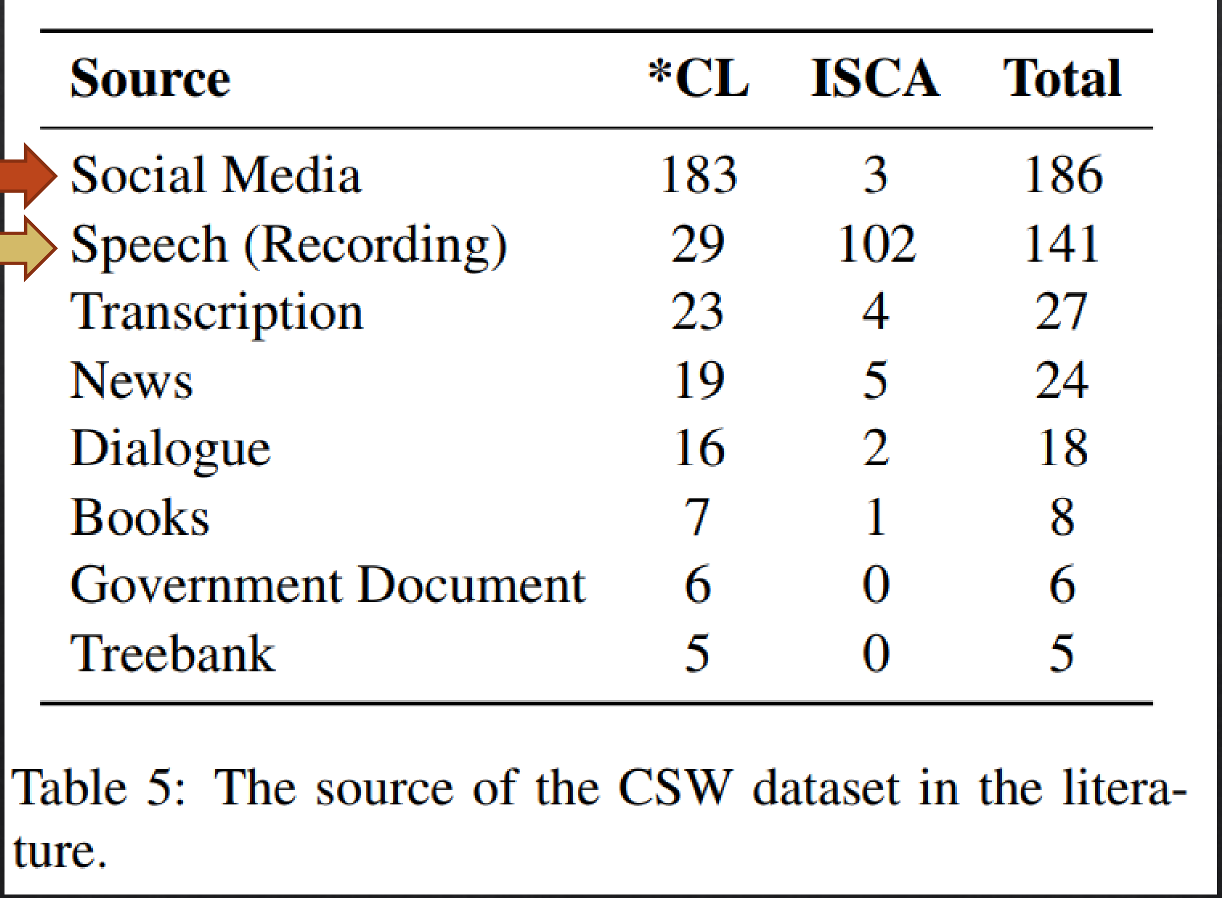
Big problem

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

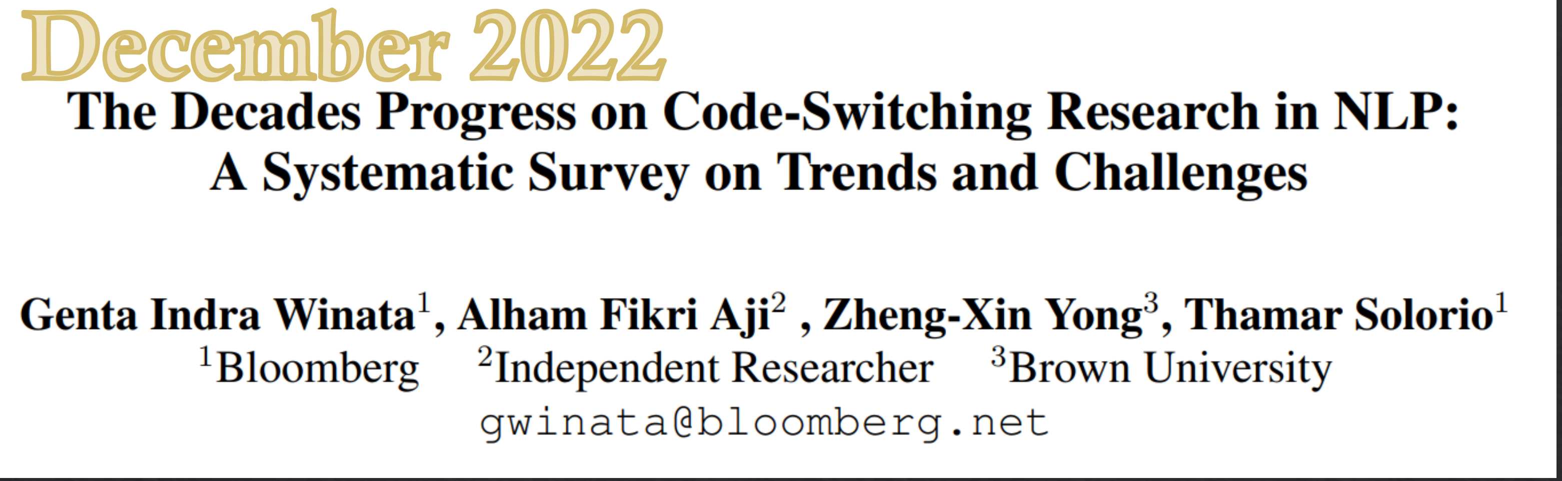
Genta Indra Winata[1], **Alham Fikri Aji**[2] , **Zheng-Xin Yong**[3], **Thamar Solorio**[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



Source	***CL**	**ISCA**	**Total**
Social Media	183	3	186
Speech (Recording)	29	102	141
Transcription	23	4	27
News	19	5	24
Dialogue	16	2	18
Books	7	1	8
Government Document	6	0	6
Treebank	5	0	5

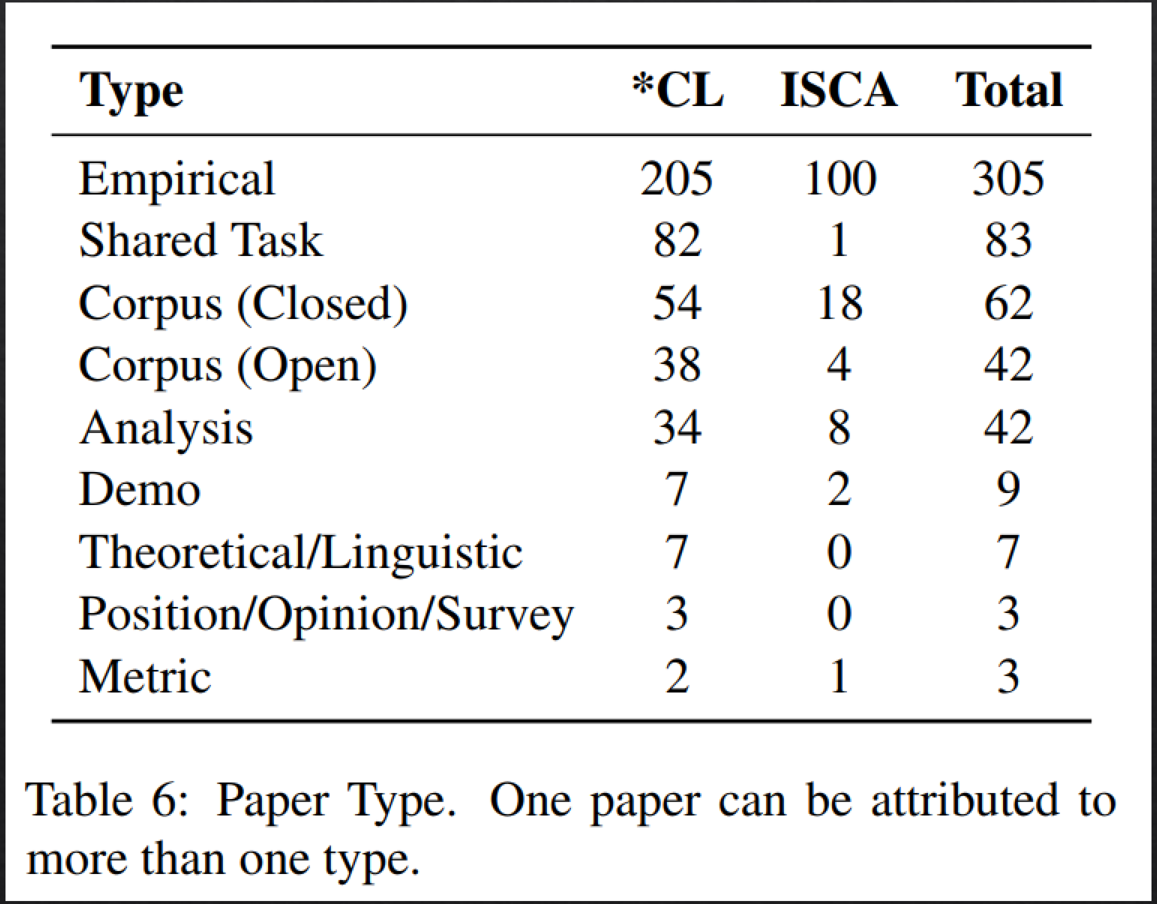
Table 5: The source of the CSW dataset in the literature.

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

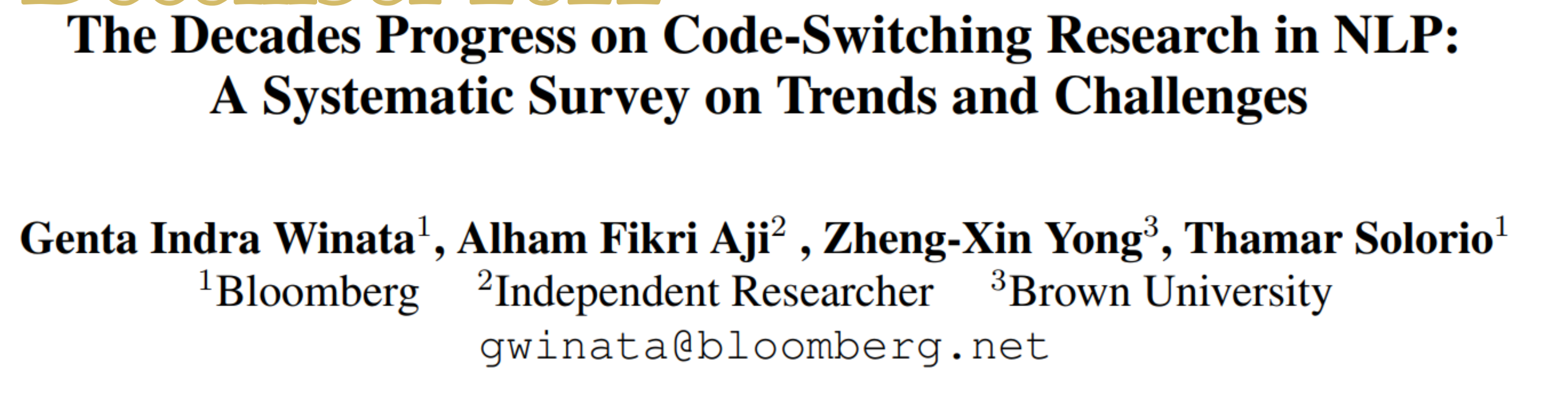
Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



Type	*CL	ISCA	Total
Empirical	205	100	305
Shared Task	82	1	83
Corpus (Closed)	54	18	62
Corpus (Open)	38	4	42
Analysis	34	8	42
Demo	7	2	9
Theoretical/Linguistic	7	0	7
Position/Opinion/Survey	3	0	3
Metric	2	1	3

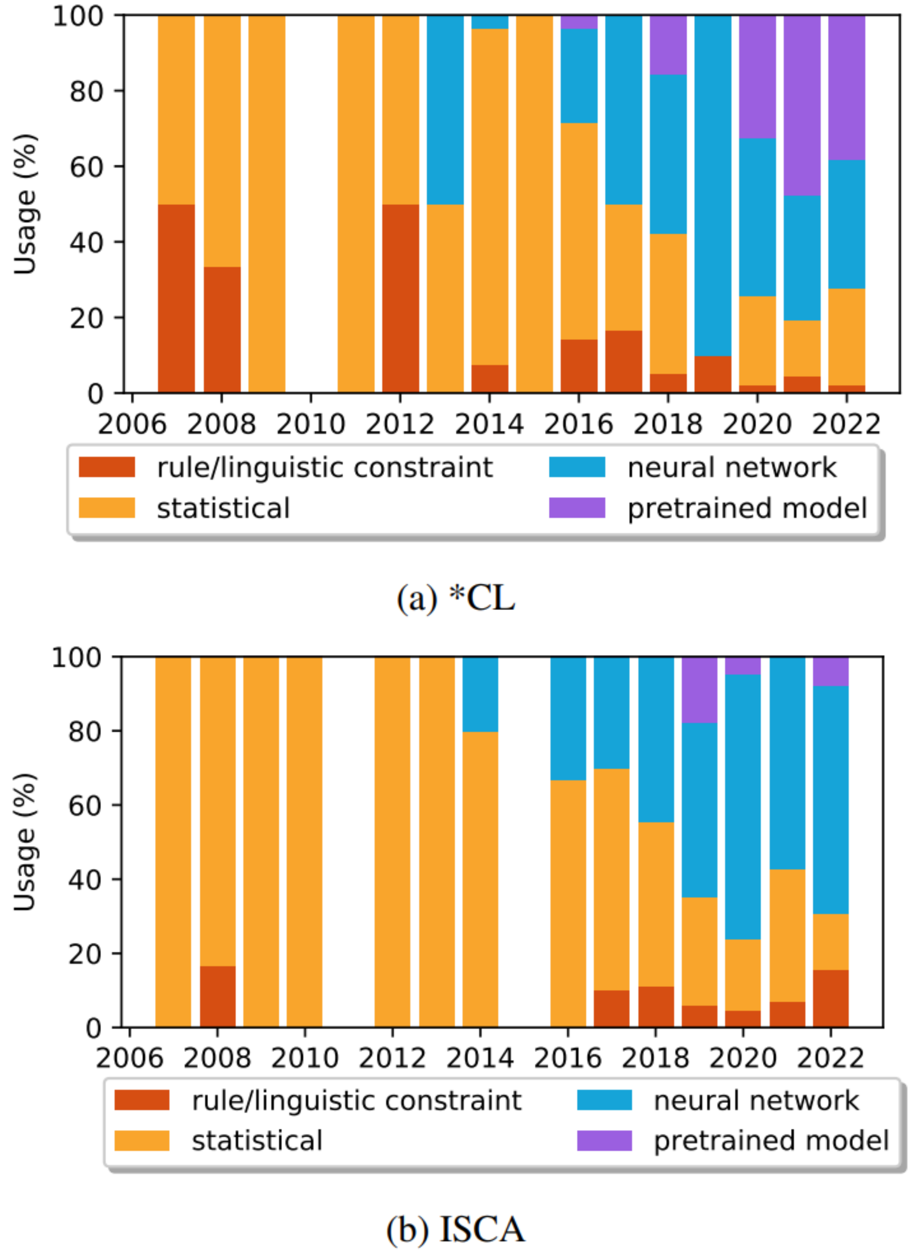
Table 6: Paper Type. One paper can be attributed to more than one type.

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

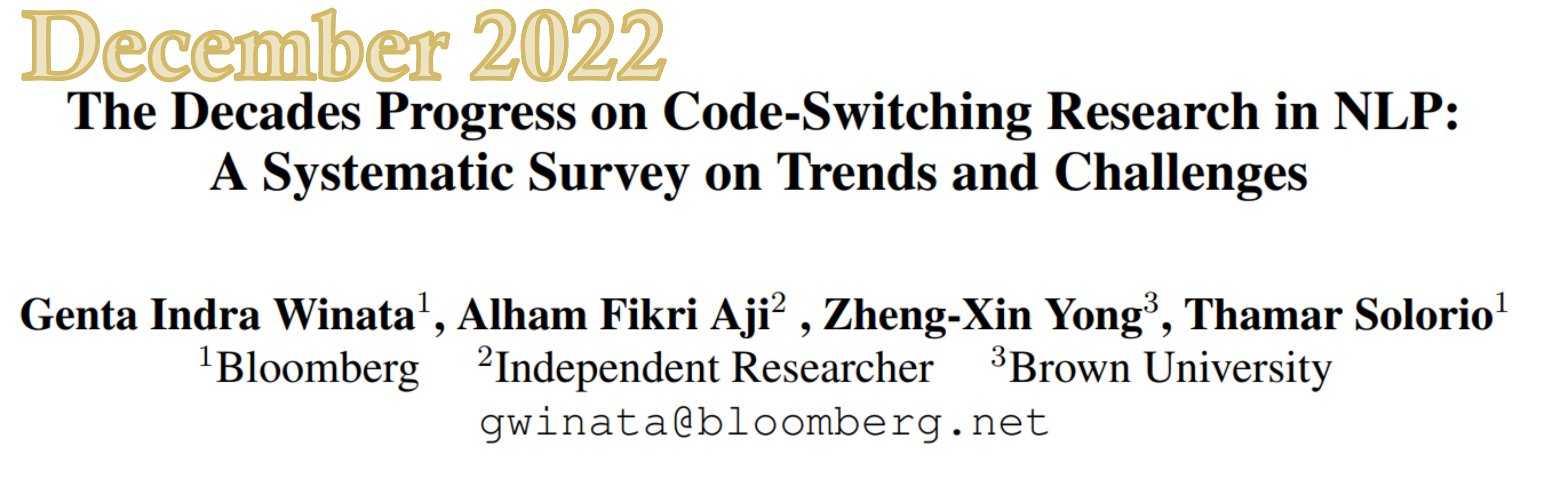


(a) *CL

(b) ISCA

Figure 5: Methods used for code-mixing NLP over the years.

December 2022

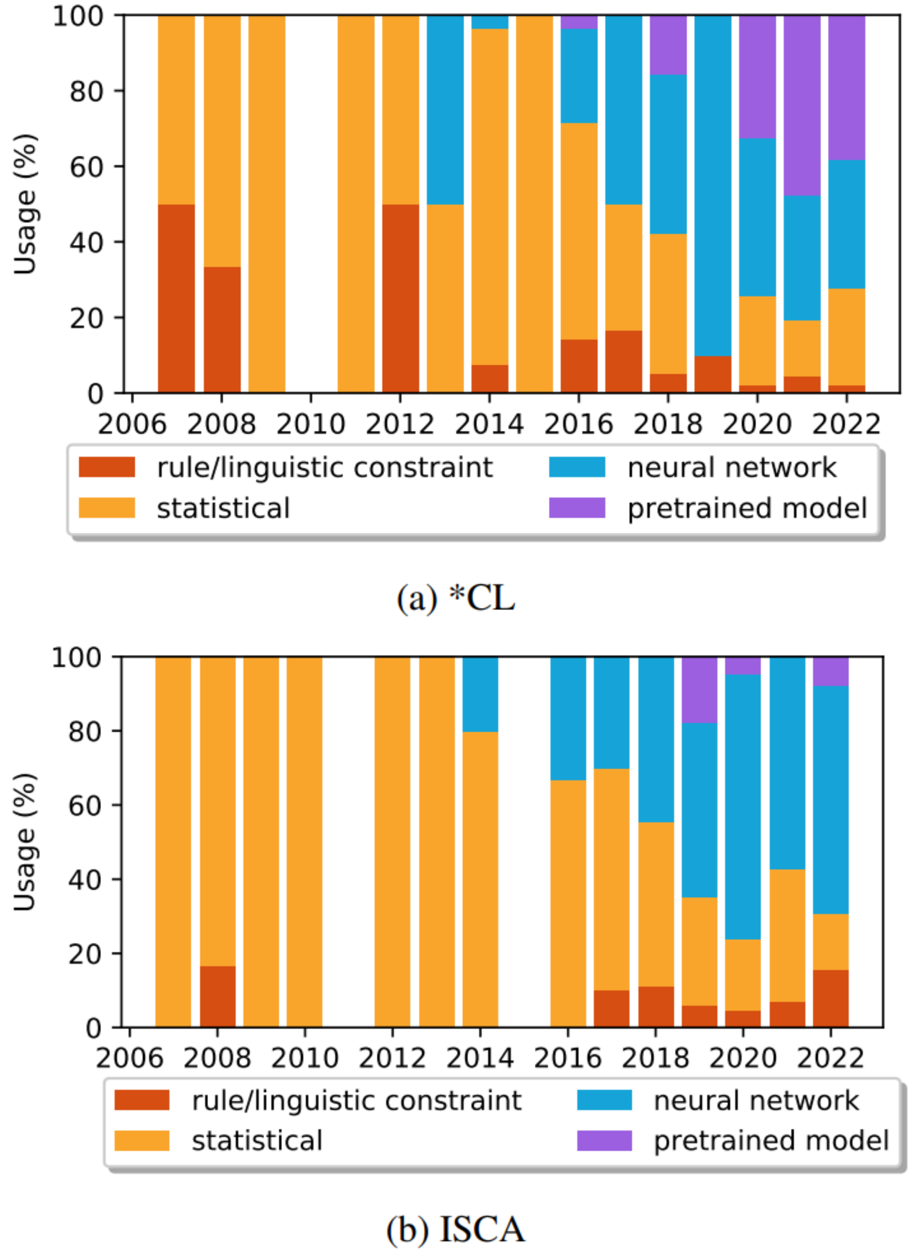


The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

Rule/Linguistic

- Equivalence Constraint
- Matrix-Embedded Language Framework
- Functional Head Constraint



(a) *CL

(b) ISCA

Figure 5: Methods used for code-mixing NLP over the years.

Equivalence Constraint

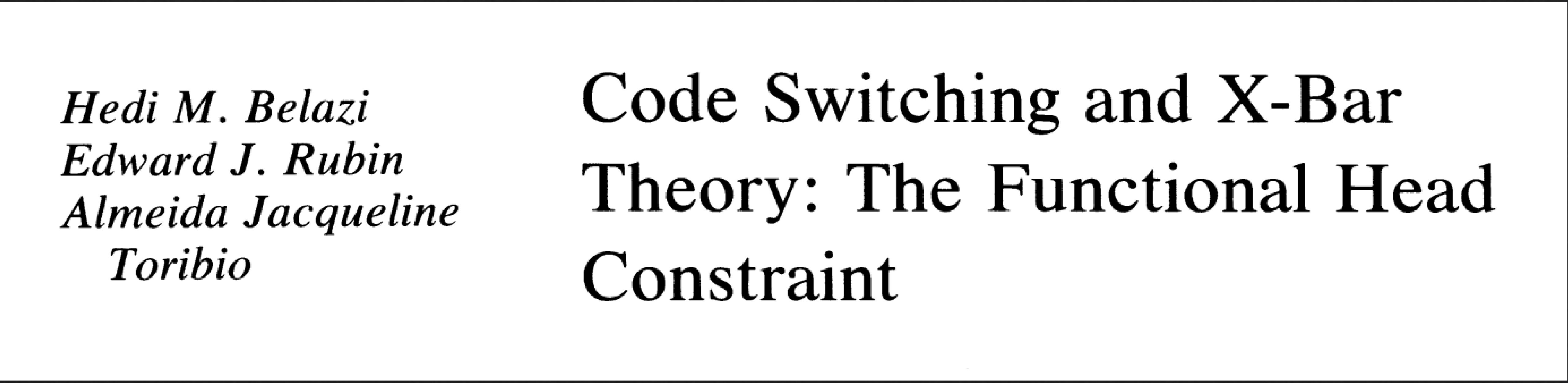
- Switching takes place where grammatical constraints of both languages satisfied (Poplack, 1980; Winata 2022)
- Parse Trees of Parallel Sentences → Match Surface order of Child Nodes (Pratapa et al 2018, 2021;Winata et al 2019)

Matrix-Embedded Language Framework (MLF)

- Asymmetrical Relationship between Languages
- Governs all or most of:
  - Grammatical Morphemes
  - Word Order
- Johnson, 1999; Myers-Scotton 1997, 2005; Lee et al., 2019; Gupta et al 2020

Functional Head Constraint

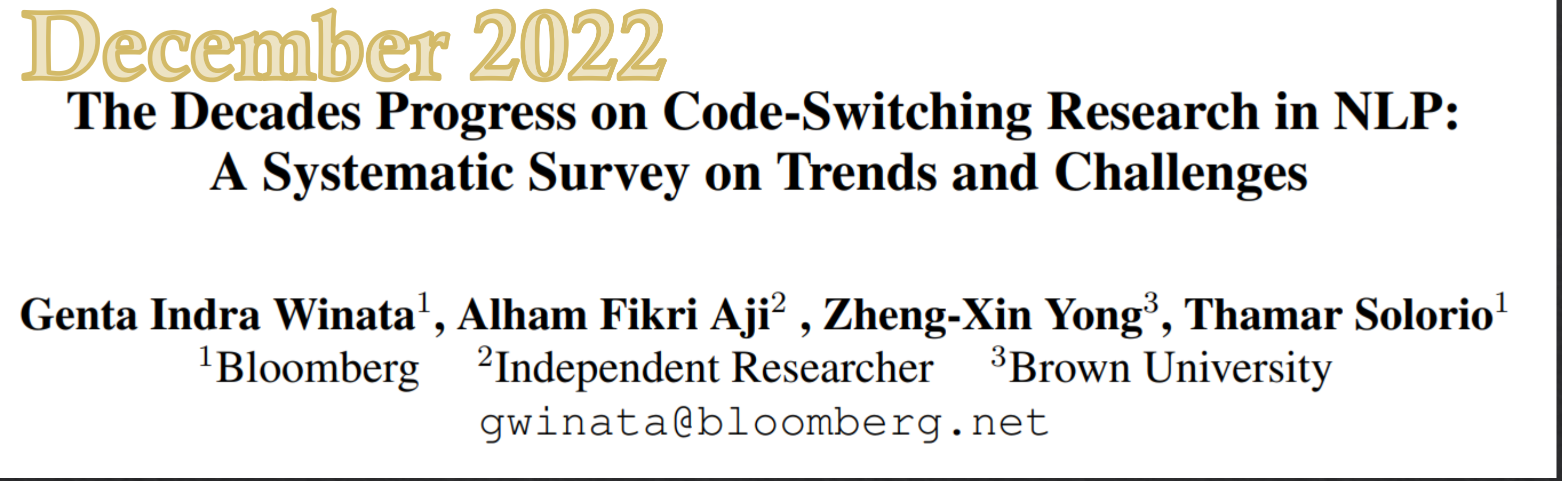
- Belazi et al., 1994
- Impossible to switch languages between functional head and its complement
- Too strong of a relationship between two constituents
- Li and Fung, 2014. Expand search in MT → Restrict path



Hedi M. Belazi
Edward J. Rubin
Almeida Jacqueline Toribio

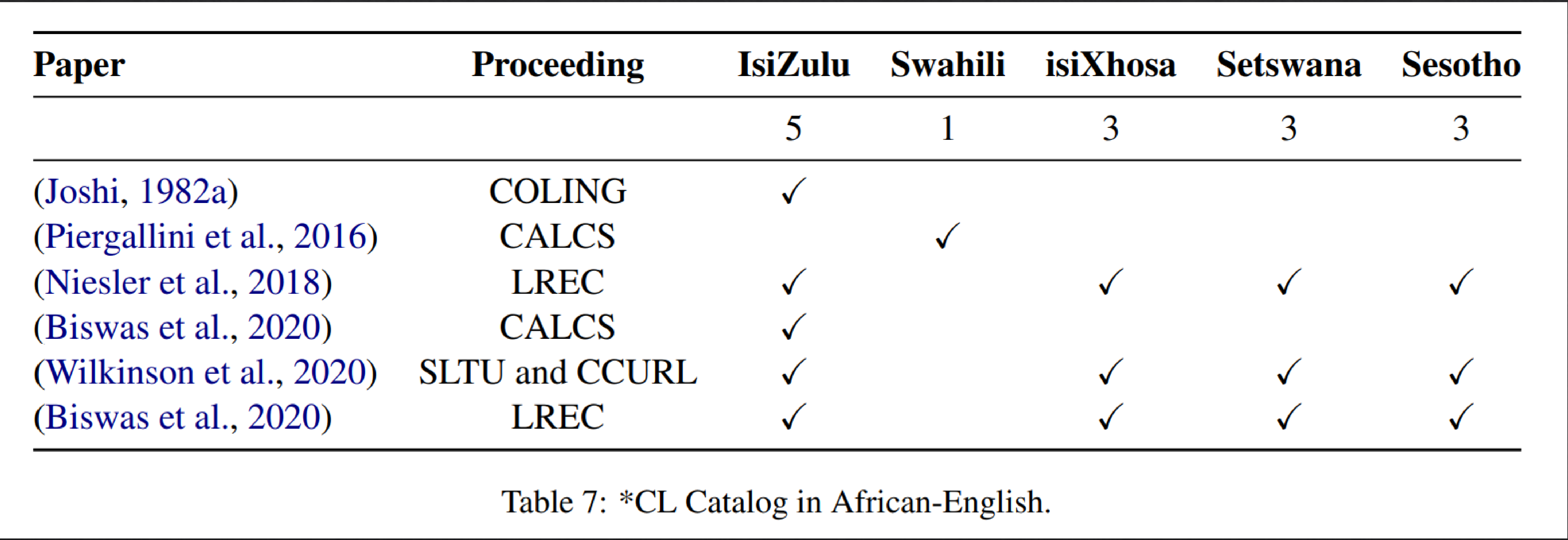
Code Switching and X-Bar Theory: The Functional Head Constraint

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

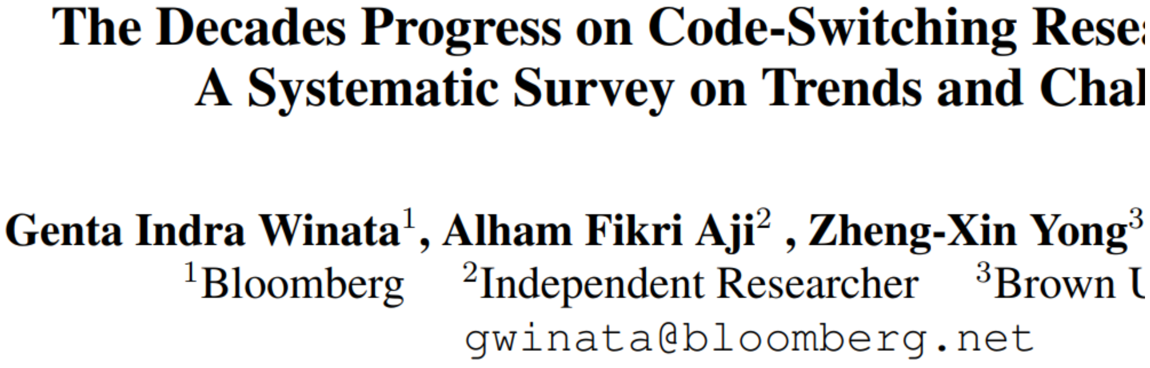
Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



Paper	Proceeding	IsiZulu	Swahili	isiXhosa	Setswana	Sesotho
		5	1	3	3	3
(Joshi, 1982a)	COLING	✓				
(Piergallini et al., 2016)	CALCS		✓			
(Niesler et al., 2018)	LREC	✓		✓	✓	✓
(Biswas et al., 2020)	CALCS	✓				
(Wilkinson et al., 2020)	SLTU and CCURL	✓		✓	✓	✓
(Biswas et al., 2020)	LREC	✓		✓	✓	✓

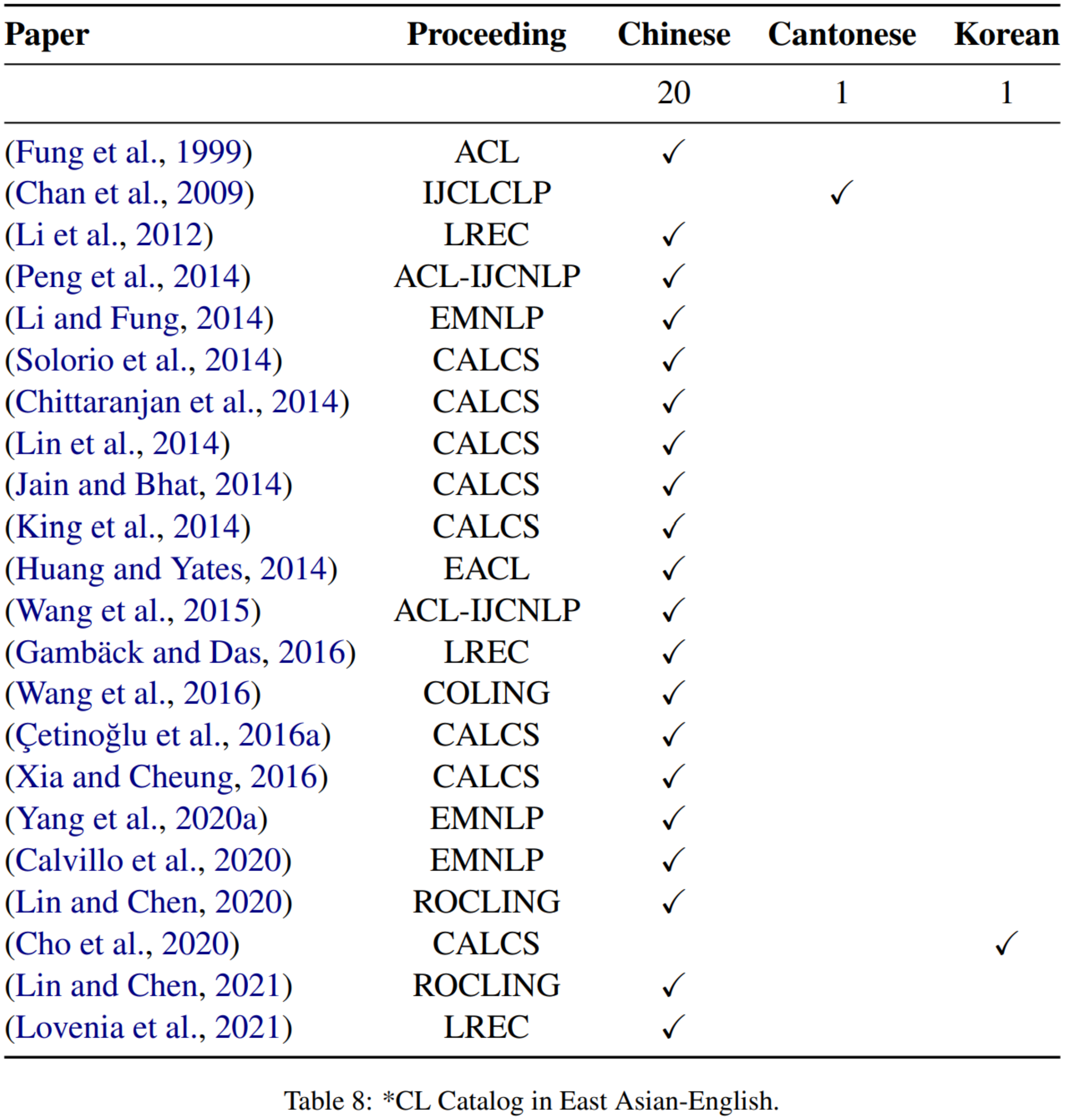
Table 7: *CL Catalog in African-English.

December 2022



The Decades Progress on Code-Switching Rese
A Systematic Survey on Trends and Chal

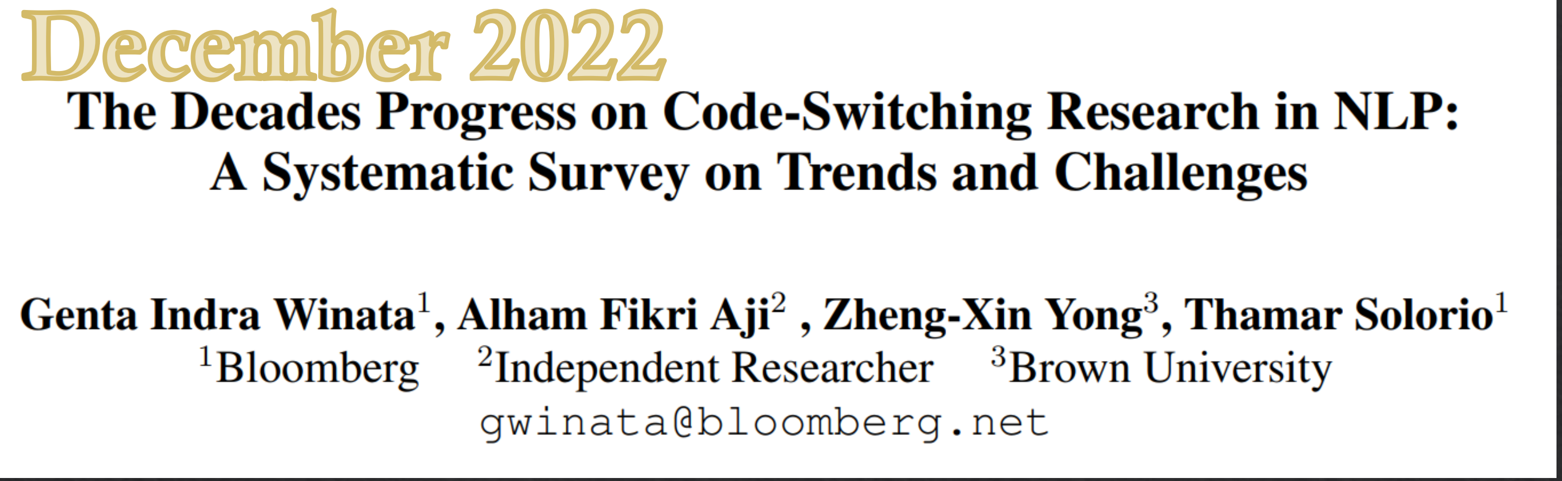
Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3]
[1]Bloomberg [2]Independent Researcher [3]Brown U
gwinata@bloomberg.net



Paper	Proceeding	Chinese	Cantonese	Korean
		20	1	1
(Fung et al., 1999)	ACL	✓		
(Chan et al., 2009)	IJCLCLP		✓	
(Li et al., 2012)	LREC	✓		
(Peng et al., 2014)	ACL-IJCNLP	✓		
(Li and Fung, 2014)	EMNLP	✓		
(Solorio et al., 2014)	CALCS	✓		
(Chittaranjan et al., 2014)	CALCS	✓		
(Lin et al., 2014)	CALCS	✓		
(Jain and Bhat, 2014)	CALCS	✓		
(King et al., 2014)	CALCS	✓		
(Huang and Yates, 2014)	EACL	✓		
(Wang et al., 2015)	ACL-IJCNLP	✓		
(Gambäck and Das, 2016)	LREC	✓		
(Wang et al., 2016)	COLING	✓		
(Çetinoğlu et al., 2016a)	CALCS	✓		
(Xia and Cheung, 2016)	CALCS	✓		
(Yang et al., 2020a)	EMNLP	✓		
(Calvillo et al., 2020)	EMNLP	✓		
(Lin and Chen, 2020)	ROCLING	✓		
(Cho et al., 2020)	CALCS			✓
(Lin and Chen, 2021)	ROCLING	✓		
(Lovenia et al., 2021)	LREC	✓		

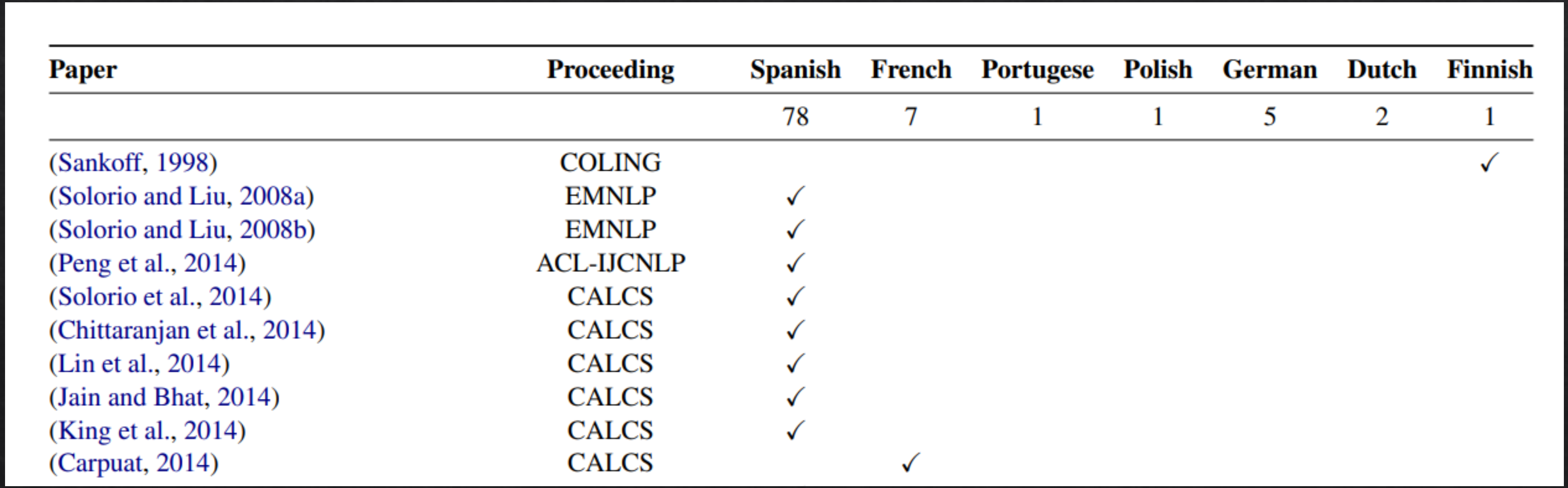
Table 8: *CL Catalog in East Asian-English.

December 2022

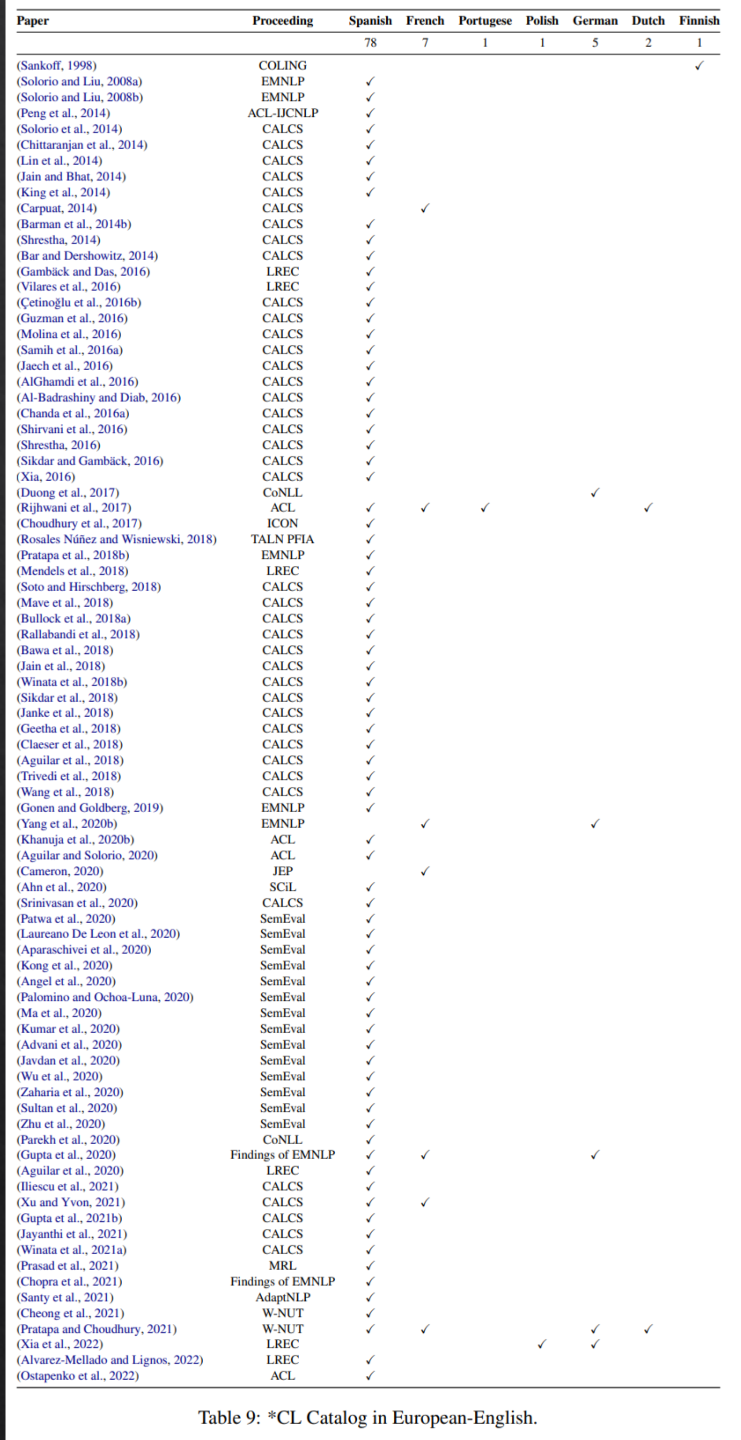


The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

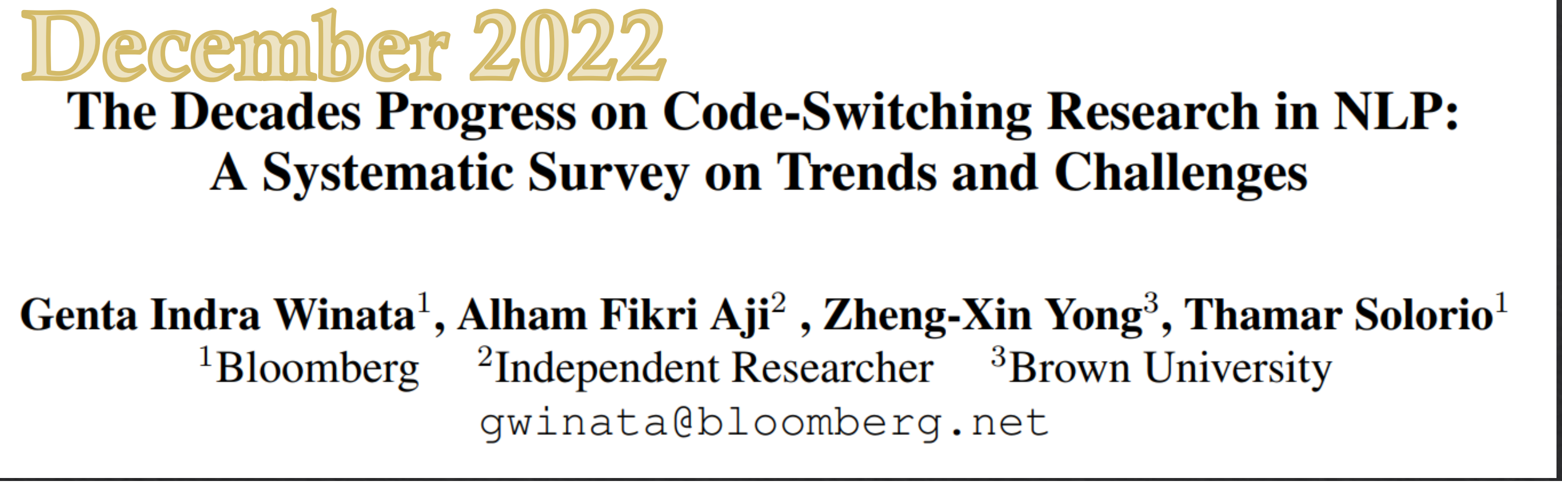


Paper	Proceeding	Spanish	French	Portugese	Polish	German	Dutch	Finnish
		78	7	1	1	5	2	1
(Sankoff, 1998)	COLING							✓
(Solorio and Liu, 2008a)	EMNLP	✓						
(Solorio and Liu, 2008b)	EMNLP	✓						
(Peng et al., 2014)	ACL-IJCNLP	✓						
(Solorio et al., 2014)	CALCS	✓						
(Chittaranjan et al., 2014)	CALCS	✓						
(Lin et al., 2014)	CALCS	✓						
(Jain and Bhat, 2014)	CALCS	✓						
(King et al., 2014)	CALCS	✓						
(Carpuat, 2014)	CALCS		✓					



Paper	Proceeding	Spanish	French	Portugese	Polish	German	Dutch	Finnish
		78	7	1	1	5	2	1
(Sankoff, 1998)	COLING							✓
(Solorio and Liu, 2008a)	EMNLP	✓						
(Solorio and Liu, 2008b)	EMNLP	✓						
(Peng et al., 2014)	ACL-IJCNLP	✓						
(Solorio et al., 2014)	CALCS	✓						
(Chittaranjan et al., 2014)	CALCS	✓						
(Lin et al., 2014)	CALCS	✓						
(Jain and Bhat, 2014)	CALCS	✓						
(King et al., 2014)	CALCS	✓						
(Carpuat, 2014)	CALCS		✓					
(Barman et al., 2014b)	CALCS	✓						
(Shrestha, 2014)	CALCS	✓						
(Bar and Dershowitz, 2014)	CALCS	✓						
(Gambäck and Das, 2016)	LREC	✓						
(Vilares et al., 2016)	LREC	✓						
(Çetinoğlu et al., 2016b)	CALCS	✓						
(Guzman et al., 2016)	CALCS	✓						
(Molina et al., 2016)	CALCS	✓						
(Samih et al., 2016a)	CALCS	✓						
(Jaech et al., 2016)	CALCS	✓						
(AlGhamdi et al., 2016)	CALCS	✓						
(Al-Badrashiny and Diab, 2016)	CALCS	✓						
(Chanda et al., 2016a)	CALCS	✓						
(Shirvani et al., 2016)	CALCS	✓						
(Shrestha, 2016)	CALCS	✓						
(Sikdar and Gambäck, 2016)	CALCS	✓						
(Xia, 2016)	CALCS	✓						
(Duong et al., 2017)	CoNLL					✓		
(Rijhwani et al., 2017)	ACL	✓	✓	✓			✓	
(Choudhury et al., 2017)	ICON	✓						
(Rosales Núñez and Wisniewski, 2018)	TALN PFIA	✓						
(Pratapa et al., 2018b)	EMNLP	✓						
(Mendels et al., 2018)	LREC	✓						
(Soto and Hirschberg, 2018)	CALCS	✓						
(Mave et al., 2018)	CALCS	✓						
(Bullock et al., 2018a)	CALCS	✓						
(Rallabandi et al., 2018)	CALCS	✓						
(Bawa et al., 2018)	CALCS	✓						
(Jain et al., 2018)	CALCS	✓						
(Winata et al., 2018b)	CALCS	✓						
(Sikdar et al., 2018)	CALCS	✓						
(Janke et al., 2018)	CALCS	✓						
(Geetha et al., 2018)	CALCS	✓						
(Claeser et al., 2018)	CALCS	✓						
(Aguilar et al., 2018)	CALCS	✓						
(Trivedi et al., 2018)	CALCS	✓						
(Wang et al., 2018)	CALCS	✓						
(Gonen and Goldberg, 2019)	EMNLP	✓						
(Yang et al., 2020b)	EMNLP		✓			✓		
(Khanuja et al., 2020b)	ACL	✓						
(Aguilar and Solorio, 2020)	ACL	✓						
(Cameron, 2020)	JEP		✓					
(Ahn et al., 2020)	SCiL	✓						
(Srinivasan et al., 2020)	CALCS	✓						
(Patwa et al., 2020)	SemEval	✓						
(Laureano De Leon et al., 2020)	SemEval	✓						
(Aparaschivei et al., 2020)	SemEval	✓						
(Kong et al., 2020)	SemEval	✓						
(Angel et al., 2020)	SemEval	✓						
(Palomino and Ochoa-Luna, 2020)	SemEval	✓						
(Ma et al., 2020)	SemEval	✓						
(Kumar et al., 2020)	SemEval	✓						
(Advani et al., 2020)	SemEval	✓						
(Javdan et al., 2020)	SemEval	✓						
(Wu et al., 2020)	SemEval	✓						
(Zaharia et al., 2020)	SemEval	✓						
(Sultan et al., 2020)	SemEval	✓						
(Zhu et al., 2020)	SemEval	✓						
(Parekh et al., 2020)	CoNLL	✓						
(Gupta et al., 2020)	Findings of EMNLP	✓	✓			✓		
(Aguilar et al., 2020)	LREC	✓						
(Iliescu et al., 2021)	CALCS	✓						
(Xu and Yvon, 2021)	CALCS	✓	✓					
(Gupta et al., 2021b)	CALCS	✓						
(Jayanthi et al., 2021)	CALCS	✓						
(Winata et al., 2021a)	CALCS	✓						
(Prasad et al., 2021)	MRL	✓						
(Chopra et al., 2021)	Findings of EMNLP	✓						
(Santy et al., 2021)	AdaptNLP	✓						
(Cheong et al., 2021)	W-NUT	✓						
(Pratapa and Choudhury, 2021)	W-NUT	✓	✓			✓	✓	
(Xia et al., 2022)	LREC				✓	✓		
(Alvarez-Mellado and Lignos, 2022)	LREC	✓						
(Ostapenko et al., 2022)	ACL	✓						

Table 9: *CL Catalog in European-English.



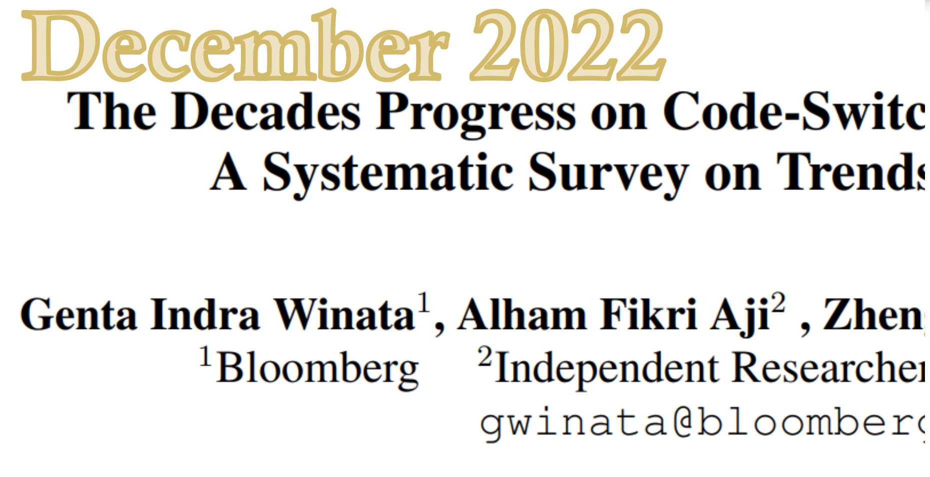
December 2022

The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

Paper	Proceeding	Egyptian Arabic	Arabic	Turkish
		3	1	2
(Rijhwani et al., 2017)	ACL			✓
(Hamed et al., 2018)	LREC	✓		
(Yirmibeşoğlu and Eryiğit, 2018)	W-NUT			✓
(Sabty et al., 2020)	WANLP		✓	
(Balabel et al., 2020)	LREC	✓		
(Hamed et al., 2020)	LREC	✓		

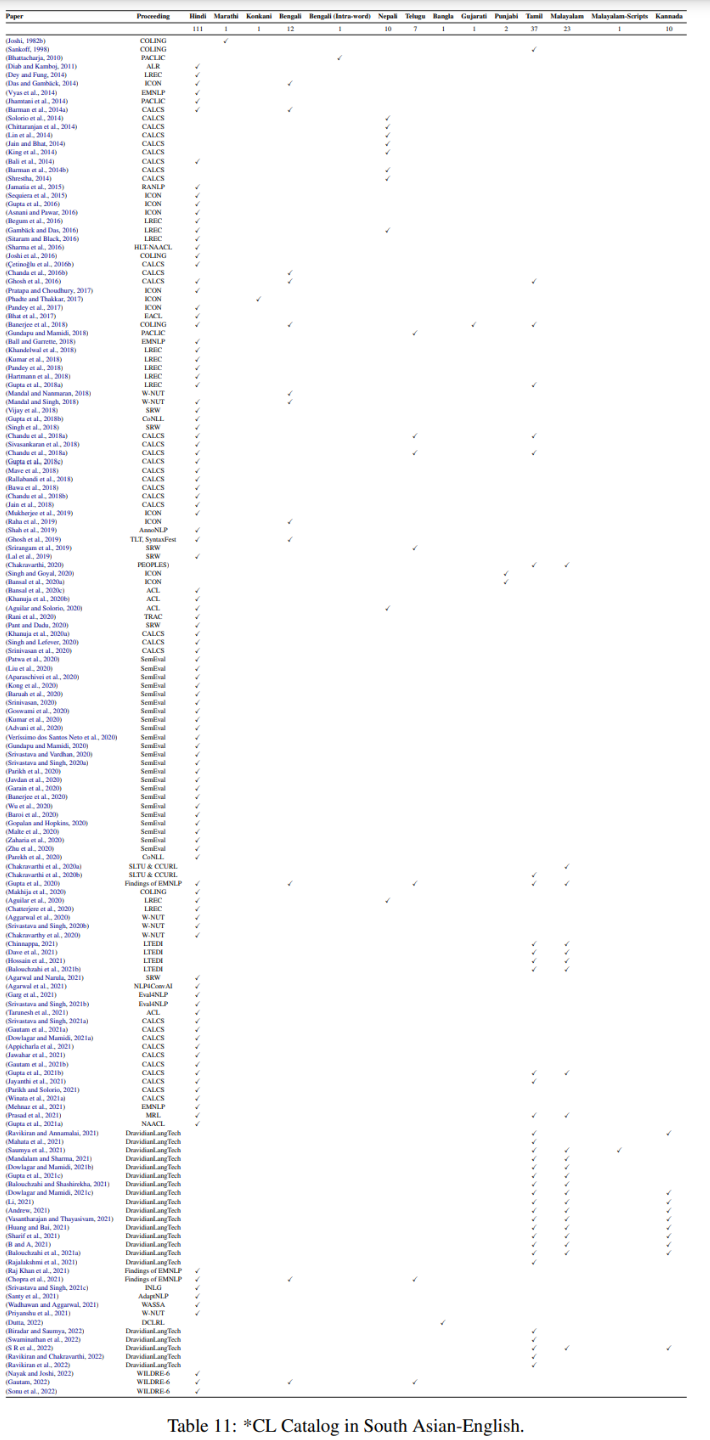
Table 10: *CL Catalog in Middle Eastern-English.



December 2022

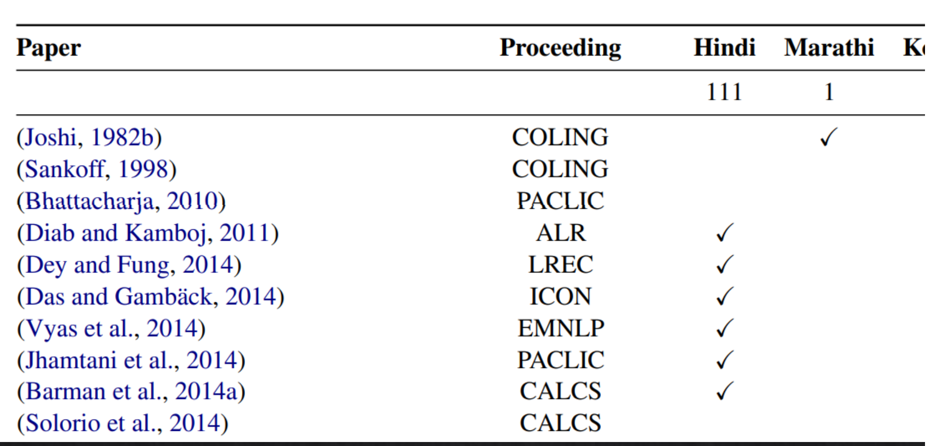
The Decades Progress on Code-Switc
A Systematic Survey on Trends

Genta Indra Winata[1], Alham Fikri Aji[2] , Zhen
[1]Bloomberg [2]Independent Researche
gwinata@bloomberg

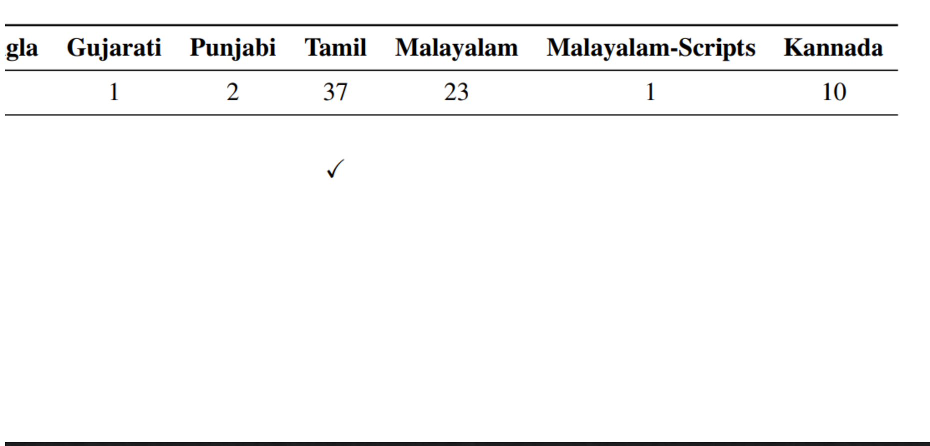


Paper	Proceeding	Hindi	Marathi	Konkani	Bengali	Bengali (Intra-word)	Nepali	Telugu	Bangla	Gujarati	Punjabi	Tamil	Malayalam	Malayalam-Scripts	Kannada
		111	1	1	12	1	10	7	1	1	2	37	23	1	10
(Joshi, 1982b)	COLING		✓												
(Sankoff, 1998)	COLING											✓			
(Bhattacharja, 2010)	PACLIC					✓									
(Diab and Kamboj, 2011)	ALR	✓													
(Dey and Fung, 2014)	LREC	✓													
(Das and Gambäck, 2014)	ICON	✓			✓										
(Vyas et al., 2014)	EMNLP	✓													
(Jhamtani et al., 2014)	PACLIC	✓													
(Barman et al., 2014a)	CALCS	✓			✓										
(Solorio et al., 2014)	CALCS						✓								
(Chittaranjan et al., 2014)	CALCS						✓								
(Lin et al., 2014)	CALCS						✓								
(Jain and Bhat, 2014)	CALCS						✓								
(King et al., 2014)	CALCS						✓								
(Bali et al., 2014)	CALCS	✓													
(Barman et al., 2014b)	CALCS						✓								
(Shrestha, 2014)	CALCS						✓								
(Jamatia et al., 2015)	RANLP	✓													
(Sequiera et al., 2015)	ICON	✓													
(Gupta et al., 2016)	ICON	✓													
(Asnani and Pawar, 2016)	ICON	✓													
(Begum et al., 2016)	LREC	✓													
(Gambäck and Das, 2016)	LREC	✓					✓								
(Sitaram and Black, 2016)	LREC	✓													
(Sharma et al., 2016)	HLT-NAACL	✓													
(Joshi et al., 2016)	COLING	✓													
(Çetinoğlu et al., 2016b)	CALCS	✓													
(Chanda et al., 2016b)	CALCS				✓										
(Ghosh et al., 2016)	CALCS	✓			✓							✓			
(Pratapa and Choudhury, 2017)	ICON	✓													
(Phadte and Thakkar, 2017)	ICON			✓											
(Pandey et al., 2017)	ICON	✓													
(Bhat et al., 2017)	EACL	✓													
(Banerjee et al., 2018)	COLING	✓			✓					✓		✓			
(Gundapu and Mamidi, 2018)	PACLIC							✓							
(Ball and Garrette, 2018)	EMNLP	✓													
(Khandelwal et al., 2018)	LREC	✓													
(Kumar et al., 2018)	LREC	✓													
(Pandey et al., 2018)	LREC	✓													
(Hartmann et al., 2018)	LREC	✓													
(Gupta et al., 2018a)	LREC	✓										✓			
(Mandal and Nanmaran, 2018)	W-NUT				✓										
(Mandal and Singh, 2018)	W-NUT	✓			✓										
(Vijay et al., 2018)	SRW	✓													
(Gupta et al., 2018b)	CoNLL	✓													
(Singh et al., 2018)	SRW	✓													
(Chandu et al., 2018a)	CALCS	✓						✓				✓			
(Sivasankaran et al., 2018)	CALCS	✓													
(Chandu et al., 2018a)	CALCS	✓						✓				✓			
(Gupta et al., 2018c)	CALCS	✓													
(Mave et al., 2018)	CALCS	✓													
(Rallabandi et al., 2018)	CALCS	✓													
(Bawa et al., 2018)	CALCS	✓													
(Chandu et al., 2018b)	CALCS	✓													
(Jain et al., 2018)	CALCS	✓													
(Mukherjee et al., 2019)	ICON	✓													
(Raha et al., 2019)	ICON				✓										
(Shah et al., 2019)	AnnoNLP	✓													
(Ghosh et al., 2019)	TLT, SyntaxFest	✓			✓										
(Srirangam et al., 2019)	SRW							✓							
(Lal et al., 2019)	SRW	✓													
(Chakravarthi, 2020)	PEOPLES)											✓	✓		
(Singh and Goyal, 2020)	ICON										✓				
(Bansal et al., 2020a)	ICON										✓				
(Bansal et al., 2020c)	ACL	✓													
(Khanuja et al., 2020b)	ACL	✓													
(Aguilar and Solorio, 2020)	ACL	✓					✓								
(Rani et al., 2020)	TRAC	✓													
(Pant and Dadu, 2020)	SRW	✓													
(Khanuja et al., 2020a)	CALCS	✓													
(Singh and Lefever, 2020)	CALCS	✓													
(Srinivasan et al., 2020)	CALCS	✓													
(Patwa et al., 2020)	SemEval	✓													
(Liu et al., 2020)	SemEval	✓													
(Aparaschivei et al., 2020)	SemEval	✓													
(Kong et al., 2020)	SemEval	✓													
(Baruah et al., 2020)	SemEval	✓													
(Srinivasan, 2020)	SemEval	✓													
(Goswami et al., 2020)	SemEval	✓													
(Kumar et al., 2020)	SemEval	✓													
(Advani et al., 2020)	SemEval	✓													
(Veríssimo dos Santos Neto et al., 2020)	SemEval	✓													
(Gundapu and Mamidi, 2020)	SemEval	✓													
(Srivastava and Vardhan, 2020)	SemEval	✓													
(Srivastava and Singh, 2020a)	SemEval	✓													
(Parikh et al., 2020)	SemEval	✓													
(Javdan et al., 2020)	SemEval	✓													
(Garain et al., 2020)	SemEval	✓													
(Banerjee et al., 2020)	SemEval	✓													
(Wu et al., 2020)	SemEval	✓													
(Baroi et al., 2020)	SemEval	✓													
(Gopalan and Hopkins, 2020)	SemEval	✓													
(Malte et al., 2020)	SemEval	✓													
(Zaharia et al., 2020)	SemEval	✓													
(Zhu et al., 2020)	SemEval	✓													
(Parekh et al., 2020)	CoNLL	✓													
(Chakravarthi et al., 2020a)	SLTU & CCURL												✓		
(Chakravarthi et al., 2020b)	SLTU & CCURL											✓			
(Gupta et al., 2020)	Findings of EMNLP	✓			✓			✓				✓	✓		
(Makhija et al., 2020)	COLING	✓													
(Aguilar et al., 2020)	LREC	✓					✓								
(Chatterjere et al., 2020)	LREC	✓													
(Aggarwal et al., 2020)	W-NUT	✓													
(Srivastava and Singh, 2020b)	W-NUT	✓													
(Chakravarthy et al., 2020)	W-NUT	✓													
(Chinnappa, 2021)	LTEDI											✓	✓		
(Dave et al., 2021)	LTEDI											✓	✓		
(Hossain et al., 2021)	LTEDI											✓	✓		
(Balouchzahi et al., 2021b)	LTEDI											✓	✓		
(Agarwal and Narula, 2021)	SRW	✓													
(Agarwal et al., 2021)	NLP4ConvAI	✓													
(Garg et al., 2021)	Eval4NLP	✓													
(Srivastava and Singh, 2021b)	Eval4NLP	✓													
(Tarunesh et al., 2021)	ACL	✓													
(Srivastava and Singh, 2021a)	CALCS	✓													
(Gautam et al., 2021a)	CALCS	✓													
(Dowlagar and Mamidi, 2021a)	CALCS	✓													
(Appicharla et al., 2021)	CALCS	✓													
(Jawahar et al., 2021)	CALCS	✓													
(Gautam et al., 2021b)	CALCS	✓													
(Gupta et al., 2021b)	CALCS	✓										✓	✓		
(Jayanthi et al., 2021)	CALCS	✓										✓			
(Parikh and Solorio, 2021)	CALCS	✓													
(Winata et al., 2021a)	CALCS	✓													
(Mehnaz et al., 2021)	EMNLP	✓													
(Prasad et al., 2021)	MRL	✓										✓	✓		
(Gupta et al., 2021a)	NAACL	✓													
(Ravikiran and Annamalai, 2021)	DravidianLangTech											✓			✓
(Mahata et al., 2021)	DravidianLangTech											✓			
(Saumya et al., 2021)	DravidianLangTech											✓	✓	✓	
(Mandalam and Sharma, 2021)	DravidianLangTech											✓	✓		
(Dowlagar and Mamidi, 2021b)	DravidianLangTech											✓	✓		
(Gupta et al., 2021c)	DravidianLangTech											✓	✓		
(Balouchzahi and Shashirekha, 2021)	DravidianLangTech											✓	✓		
(Dowlagar and Mamidi, 2021c)	DravidianLangTech											✓	✓		✓
(Li, 2021)	DravidianLangTech											✓	✓		✓
(Andrew, 2021)	DravidianLangTech											✓	✓		✓
(Vasantharajan and Thayasivam, 2021)	DravidianLangTech											✓	✓		✓
(Huang and Bai, 2021)	DravidianLangTech											✓	✓		✓
(Sharif et al., 2021)	DravidianLangTech											✓	✓		✓
(B and A, 2021)	DravidianLangTech											✓	✓		✓
(Balouchzahi et al., 2021a)	DravidianLangTech											✓	✓		✓
(Rajalakshmi et al., 2021)	DravidianLangTech											✓			
(Raj Khan et al., 2021)	Findings of EMNLP	✓													
(Chopra et al., 2021)	Findings of EMNLP	✓			✓			✓							
(Srivastava and Singh, 2021c)	INLG	✓													
(Santy et al., 2021)	AdaptNLP	✓													
(Wadhawan and Aggarwal, 2021)	WASSA	✓													
(Priyanshu et al., 2021)	W-NUT	✓													
(Dutta, 2022)	DCLRL								✓						
(Biradar and Saumya, 2022)	DravidianLangTech											✓			
(Swaminathan et al., 2022)	DravidianLangTech											✓			
(S R et al., 2022)	DravidianLangTech											✓	✓		✓
(Ravikiran and Chakravarthi, 2022)	DravidianLangTech											✓			
(Ravikiran et al., 2022)	DravidianLangTech											✓			
(Nayak and Joshi, 2022)	WILDRE-6	✓													
(Gautam, 2022)	WILDRE-6	✓			✓			✓							
(Sonu et al., 2022)	WILDRE-6	✓													

Table 11: *CL Catalog in South Asian-English.

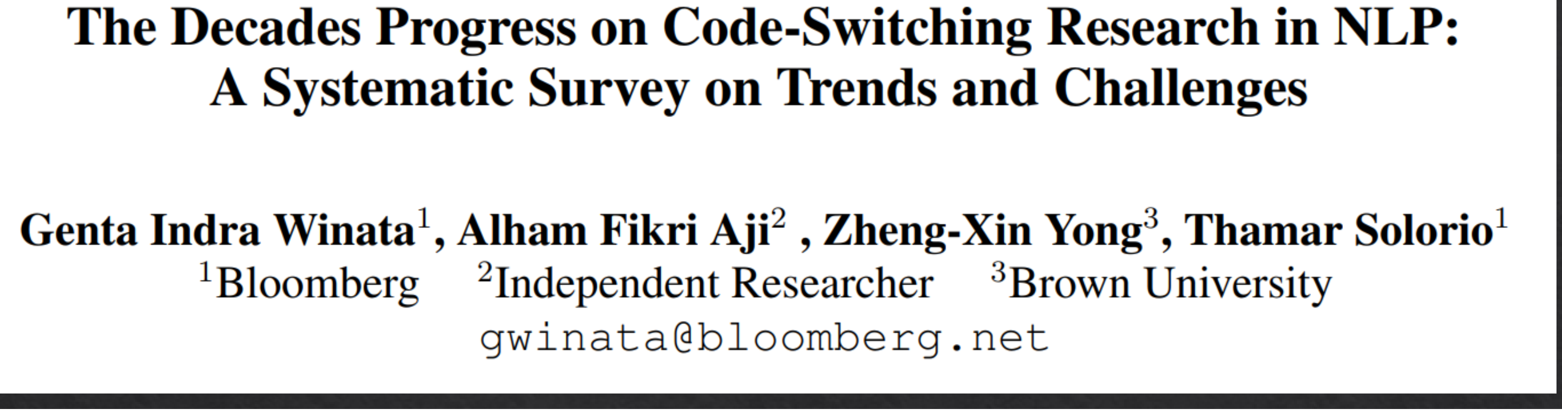


Paper	Proceeding	Hindi	Marathi	K
		111	1	
(Joshi, 1982b)	COLING		✓	
(Sankoff, 1998)	COLING			
(Bhattacharja, 2010)	PACLIC			
(Diab and Kamboj, 2011)	ALR	✓		
(Dey and Fung, 2014)	LREC	✓		
(Das and Gambäck, 2014)	ICON	✓		
(Vyas et al., 2014)	EMNLP	✓		
(Jhamtani et al., 2014)	PACLIC	✓		
(Barman et al., 2014a)	CALCS	✓		
(Solorio et al., 2014)	CALCS			



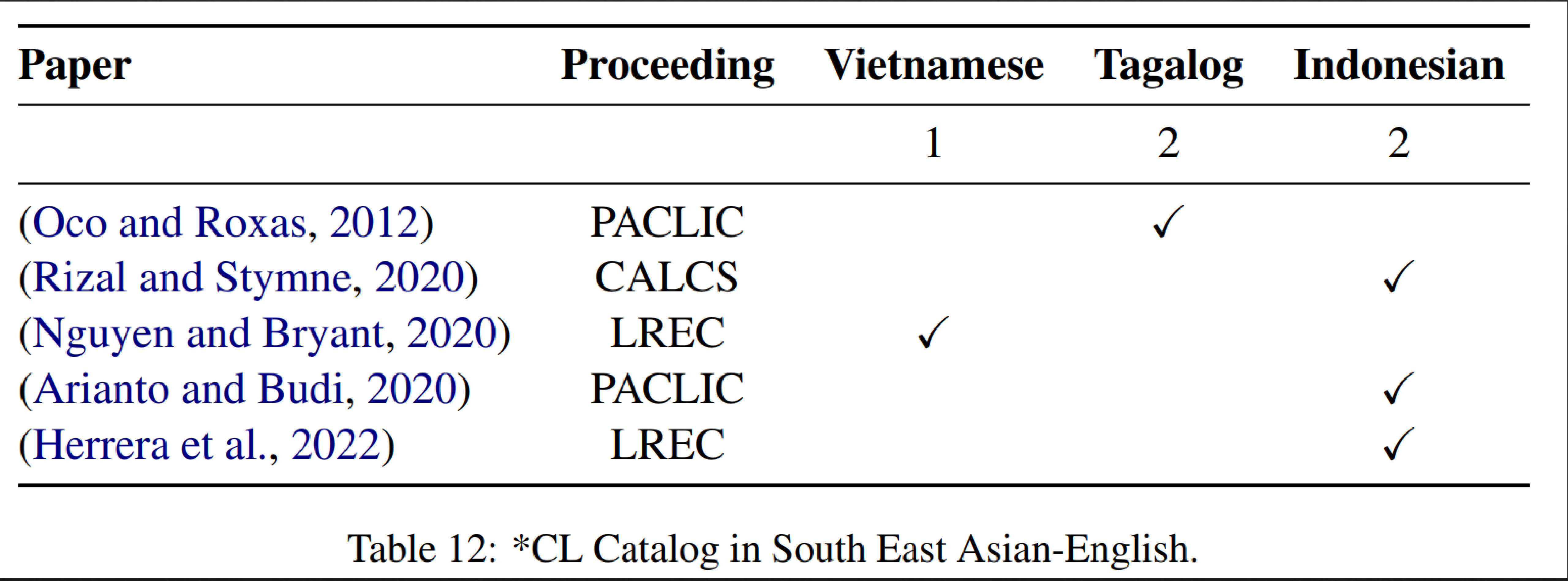
gla	Gujarati	Punjabi	Tamil	Malayalam	Malayalam-Scripts	Kannada
	1	2	37	23	1	10
			✓			

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



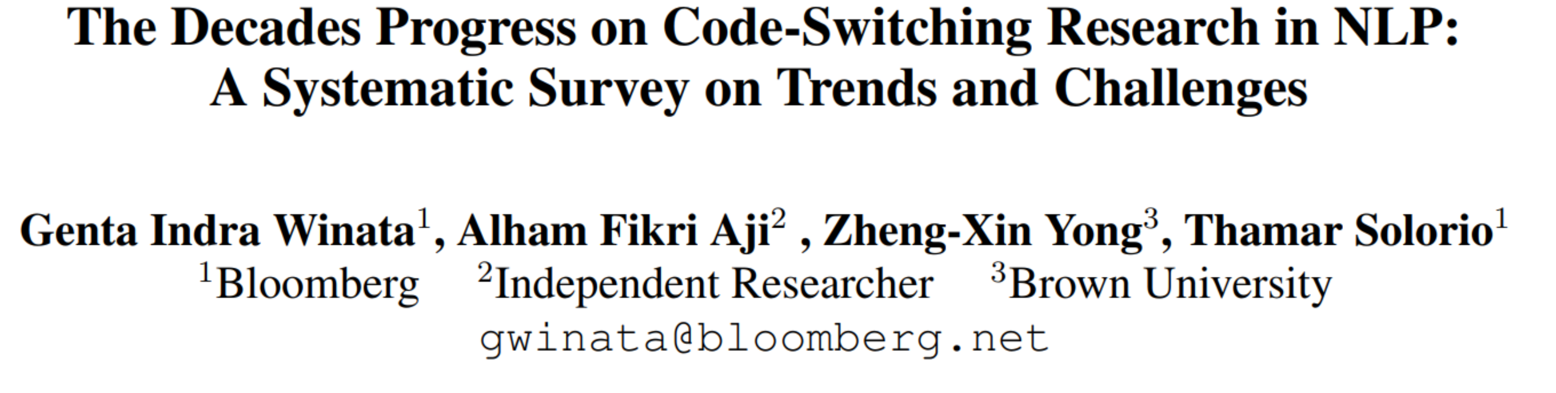
Paper	Proceeding	Vietnamese	Tagalog	Indonesian
		1	2	2
(Oco and Roxas, 2012)	PACLIC		✓	
(Rizal and Stymne, 2020)	CALCS			✓
(Nguyen and Bryant, 2020)	LREC	✓		
(Arianto and Budi, 2020)	PACLIC			✓
(Herrera et al., 2022)	LREC			✓

Table 12: *CL Catalog in South East Asian-English.

Paper	Proceeding	Darija-MSA	MSA-Egyptian	MSA-Other Dialect	Chinese-Taiwanese	MSA-Levant Arabic	MSA-Gulf	Mixed-English
		1	15	10	2	2	1	1
(Chu et al., 2007)				✓				
(Yu et al., 2012)	CIPS-SIGHAN				✓			
(Elfardy and Diab, 2012)	COLING		✓			✓		
(Solorio et al., 2014)	CALCS			✓				
(Chittaranjan et al., 2014)	CALCS			✓				
(Lin et al., 2014)	CALCS			✓				
(Jain and Bhat, 2014)	CALCS			✓				
(Elfardy et al., 2014)	CALCS			✓				
(King et al., 2014)	CALCS			✓				
(Gambäck and Das, 2016)	LREC		✓					
(Samih and Maier, 2016)	LREC	✓						
(Diab et al., 2016)	LREC		✓					
(Molina et al., 2016)	CALCS		✓					
(Samih et al., 2016a)	CALCS		✓					
(Jaech et al., 2016)	CALCS			✓				
(Samih et al., 2016b)	CALCS			✓				
(AlGhamdi et al., 2016)	CALCS		✓					
(Al-Badrashiny and Diab, 2016)	CALCS			✓				
(Shrestha, 2016)	CALCS			✓				
(El-Haj et al., 2018)	LREC		✓			✓	✓	
(Shoemark et al., 2018)	W-NUT							✓
(Attia et al., 2018)	CALCS		✓					
(Janke et al., 2018)	CALCS		✓					
(Geetha et al., 2018)	CALCS		✓					
(Aguilar et al., 2018)	CALCS		✓					
(Wang et al., 2018)	CALCS		✓					
(Aguilar et al., 2020)	LREC		✓					
(Nagoudi et al., 2021)	CALCS		✓					
(Winata et al., 2021a)	CALCS		✓					

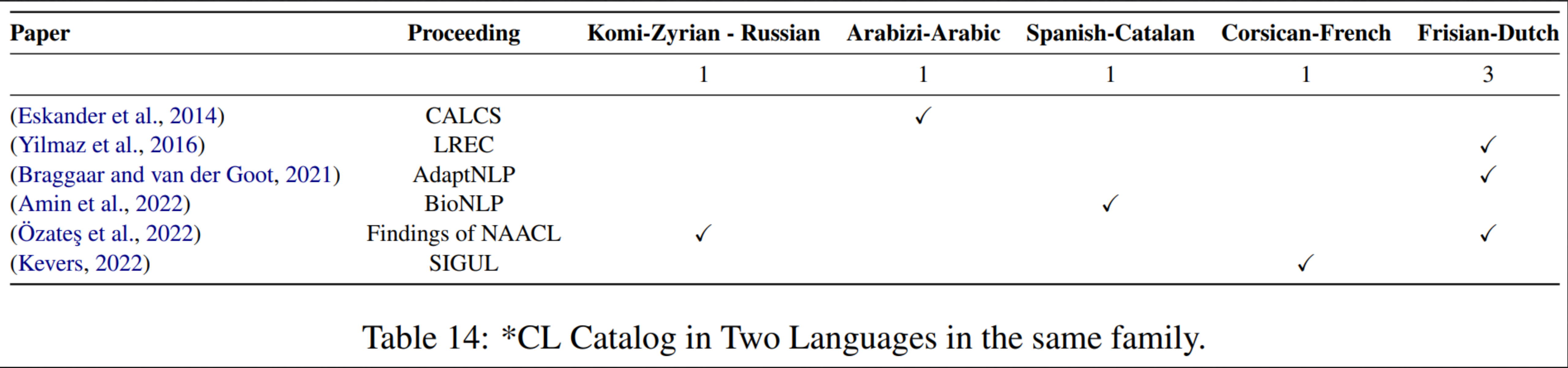
Table 13: *CL Catalog in Dialect.

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], **Alham Fikri Aji**[2] , **Zheng-Xin Yong**[3], **Thamar Solorio**[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



Paper	**Proceeding**	**Komi-Zyrian - Russian**	**Arabizi-Arabic**	**Spanish-Catalan**	**Corsican-French**	**Frisian-Dutch**
		1	1	1	1	3
(Eskander et al., 2014)	CALCS		✓			
(Yilmaz et al., 2016)	LREC					✓
(Braggaar and van der Goot, 2021)	AdaptNLP					✓
(Amin et al., 2022)	BioNLP			✓		
(Özateş et al., 2022)	Findings of NAACL	✓				✓
(Kevers, 2022)	SIGUL				✓	

Table 14: *CL Catalog in Two Languages in the same family.

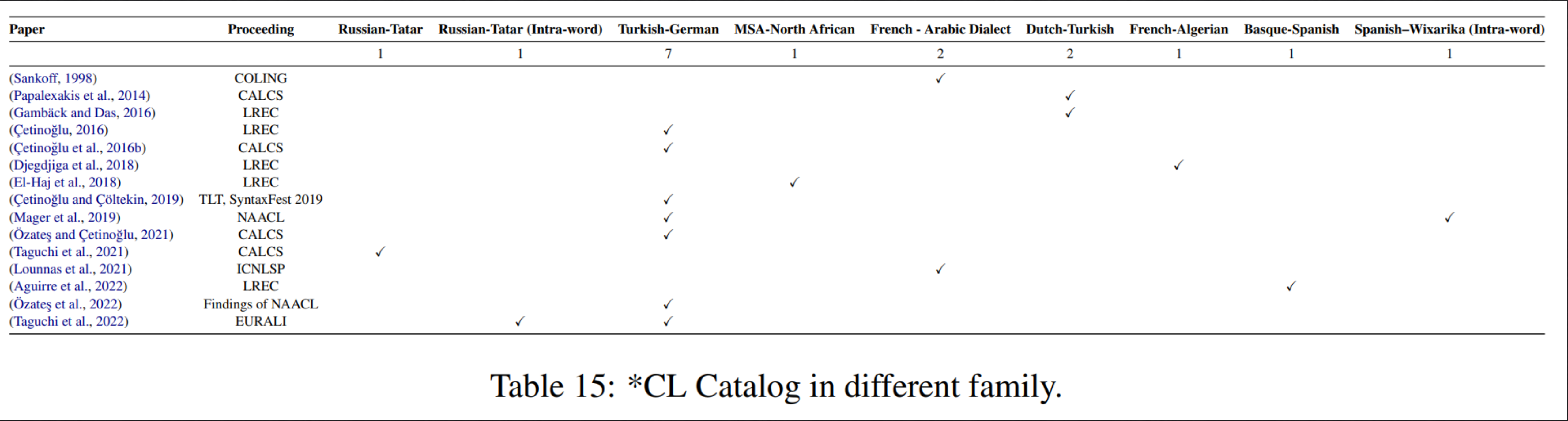
Hmmmm …. Family?

Spanglish? Dialectal

December 2022

The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

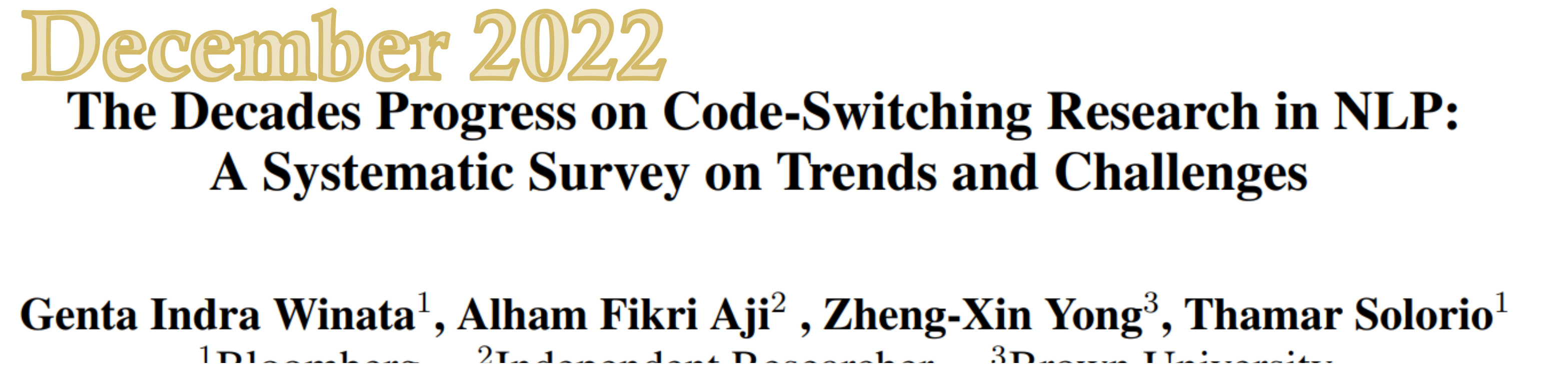


Paper	Proceeding	Russian-Tatar	Russian-Tatar (Intra-word)	Turkish-German	MSA-North African	French - Arabic Dialect	Dutch-Turkish	French-Algerian	Basque-Spanish	Spanish–Wixarika (Intra-word)
		1	1	7	1	2	2	1	1	1
(Sankoff, 1998)	COLING					✓				
(Papalexakis et al., 2014)	CALCS						✓			
(Gambäck and Das, 2016)	LREC						✓			
(Çetinoğlu, 2016)	LREC			✓						
(Çetinoğlu et al., 2016b)	CALCS			✓						
(Djegdjiga et al., 2018)	LREC							✓		
(El-Haj et al., 2018)	LREC				✓					
(Çetinoğlu and Çöltekin, 2019)	TLT, SyntaxFest 2019			✓						
(Mager et al., 2019)	NAACL			✓						✓
(Özateş and Çetinoğlu, 2021)	CALCS			✓						
(Taguchi et al., 2021)	CALCS	✓								
(Lounnas et al., 2021)	ICNLSP					✓				
(Aguirre et al., 2022)	LREC								✓	
(Özateş et al., 2022)	Findings of NAACL			✓						
(Taguchi et al., 2022)	EURALI		✓	✓						

Table 15: *CL Catalog in different family.

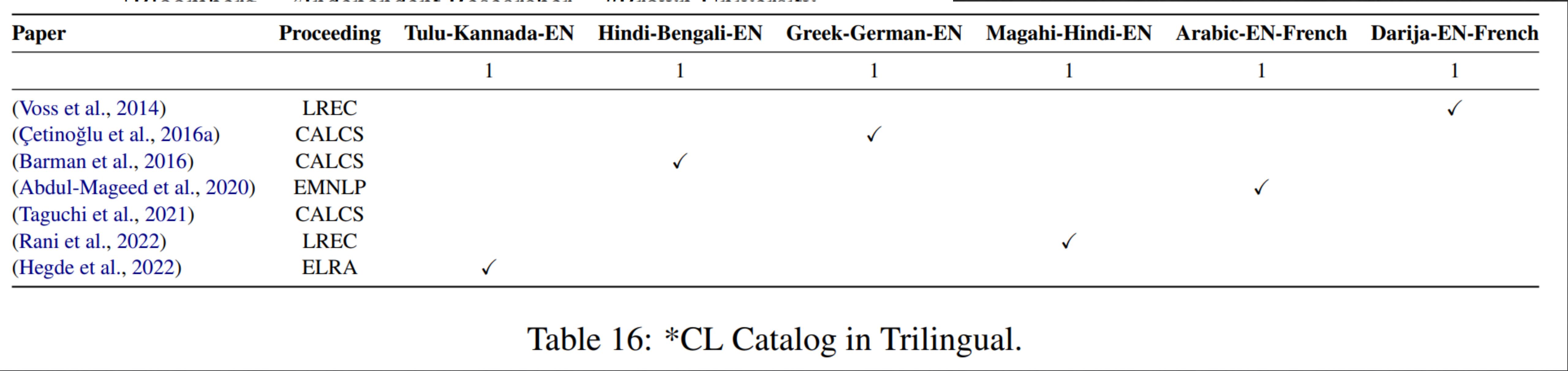
Ok, Agree Here

December 2022



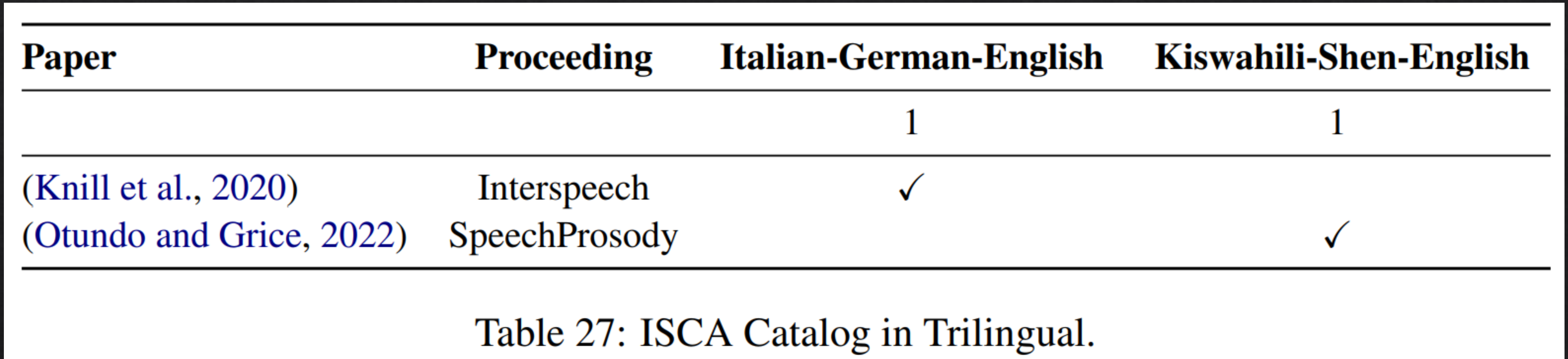
The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University



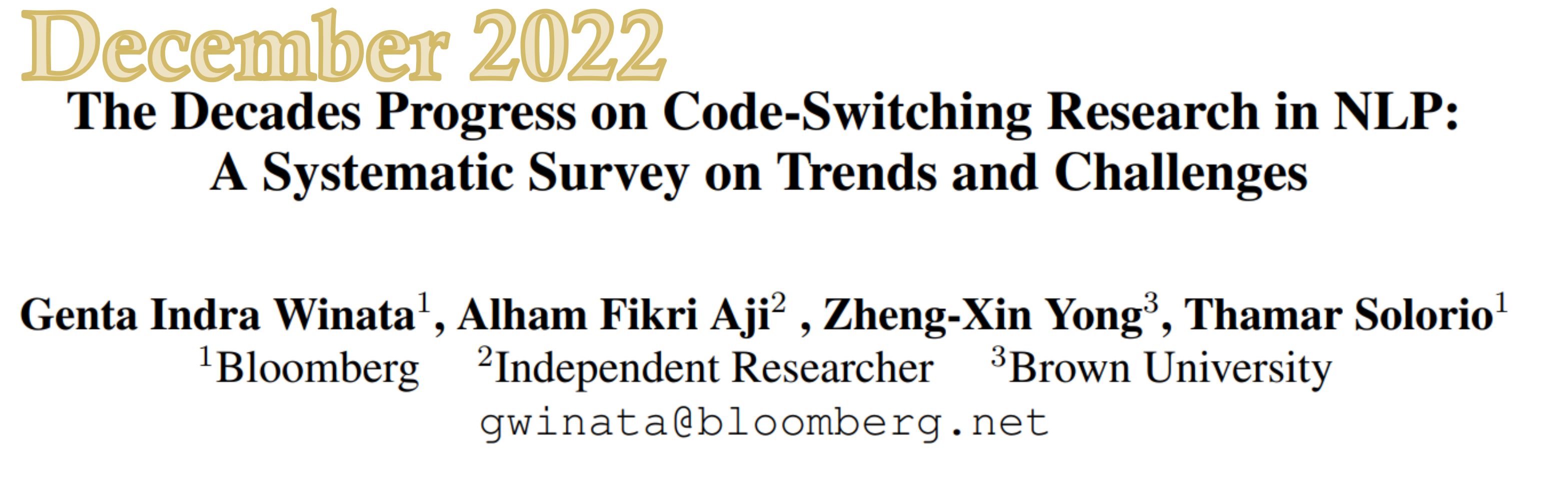
Paper	Proceeding	Tulu-Kannada-EN	Hindi-Bengali-EN	Greek-German-EN	Magahi-Hindi-EN	Arabic-EN-French	Darija-EN-French
		1	1	1	1	1	1
(Voss et al., 2014)	LREC						✓
(Çetinoğlu et al., 2016a)	CALCS			✓			
(Barman et al., 2016)	CALCS		✓				
(Abdul-Mageed et al., 2020)	EMNLP					✓	
(Taguchi et al., 2021)	CALCS						
(Rani et al., 2022)	LREC				✓		
(Hegde et al., 2022)	ELRA	✓					

Table 16: *CL Catalog in Trilingual.



Paper	Proceeding	Italian-German-English	Kiswahili-Shen-English
		1	1
(Knill et al., 2020)	Interspeech	✓	
(Otundo and Grice, 2022)	SpeechProsody		✓

Table 27: ISCA Catalog in Trilingual.

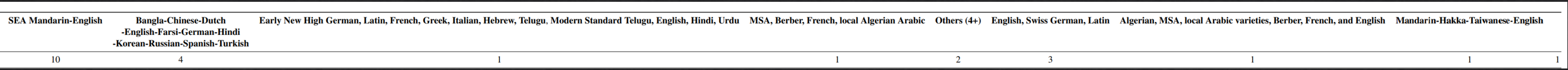


December 2022

The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

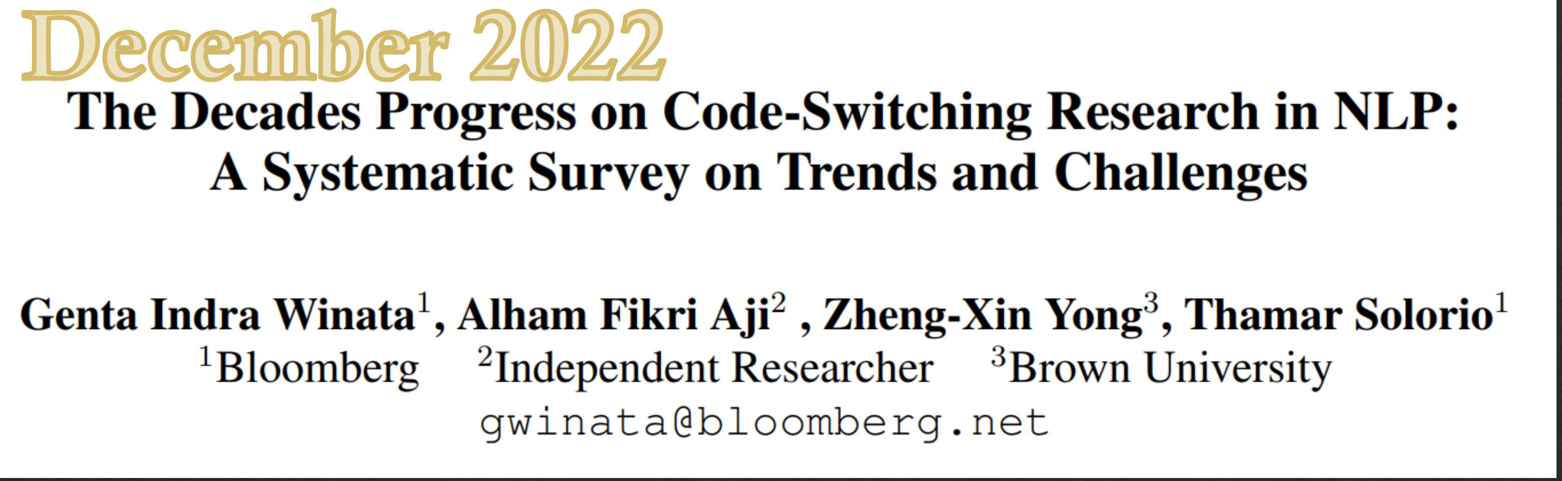
Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net

- 4+ Languages



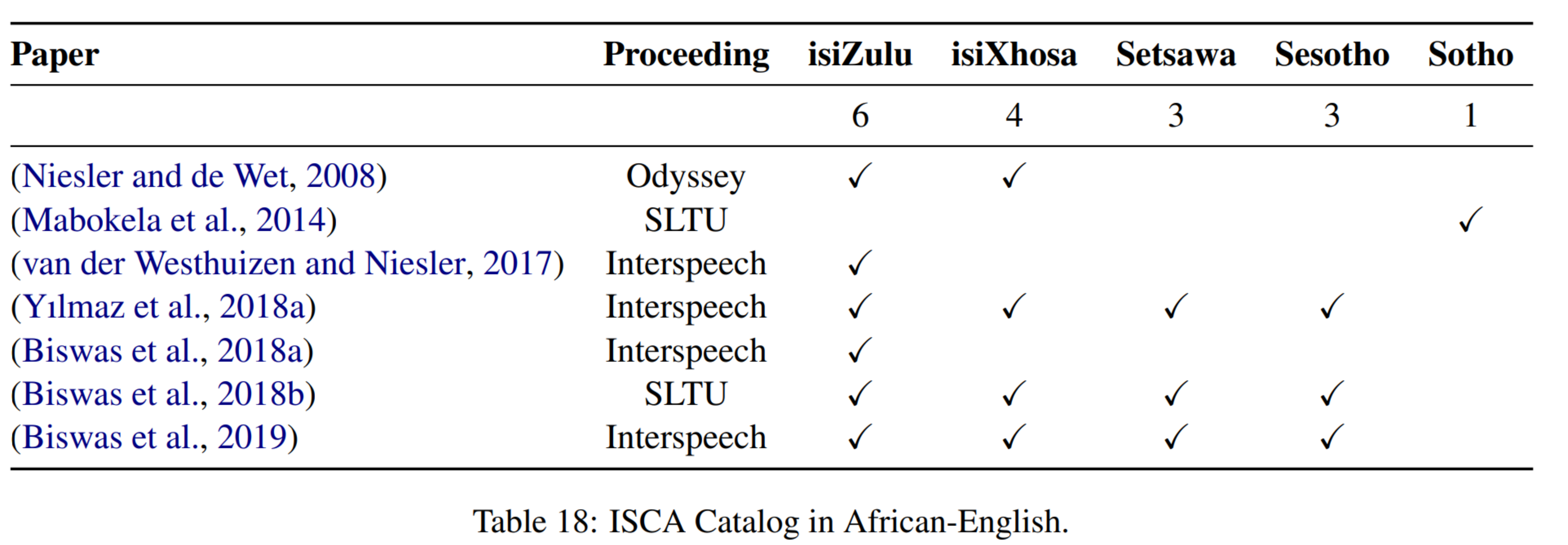
SEA Mandarin-English	Bangla-Chinese-Dutch -English-Farsi-German-Hindi -Korean-Russian-Spanish-Turkish	Early New High German, Latin, French, Greek, Italian, Hebrew, Telugu, Modern Standard Telugu, English, Hindi, Urdu	MSA, Berber, French, local Algerian Arabic	Others (4+)	English, Swiss German, Latin	Algerian, MSA, local Arabic varieties, Berber, French, and English	Mandarin-Hakka-Taiwanese-English	
10	4	1	1	2	3	1	1	1

December 2022



The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



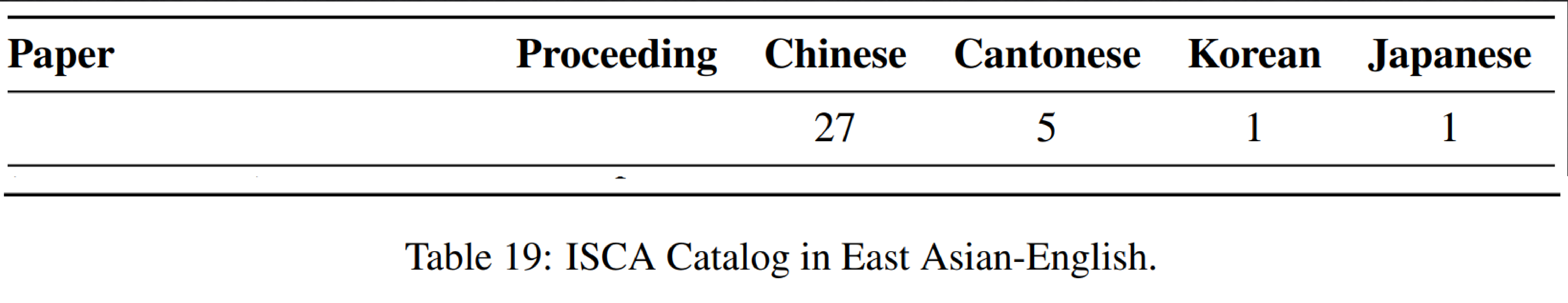
Paper	**Proceeding**	**isiZulu**	**isiXhosa**	**Setsawa**	**Sesotho**	**Sotho**
		6	4	3	3	1
(Niesler and de Wet, 2008)	Odyssey	✓	✓			
(Mabokela et al., 2014)	SLTU					✓
(van der Westhuizen and Niesler, 2017)	Interspeech	✓				
(Yılmaz et al., 2018a)	Interspeech	✓	✓	✓	✓	
(Biswas et al., 2018a)	Interspeech	✓				
(Biswas et al., 2018b)	SLTU	✓	✓	✓	✓	
(Biswas et al., 2019)	Interspeech	✓	✓	✓	✓	

Table 18: ISCA Catalog in African-English.

December 2022

The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], Alham Fikri Aji[2] , Zheng-Xin Yong[3], Thamar Solorio[1]
[1]Bloomberg [2]Independent Researcher [3]Brown University
gwinata@bloomberg.net



Paper	Proceeding	Chinese	Cantonese	Korean	Japanese
		27	5	1	1

Table 19: ISCA Catalog in East Asian-English.

Paper	Proceeding	Spanish	French	German	Maltese
(Pfister and Romsdorfer, 2003)	Eurospeech			✓	
(Romsdorfer and Pfister, 2005)	Interspeech		✓		
(Rosner and Farrugia, 2007)	Interspeech				✓
(Piccinini and Garellek, 2014)	SpeechProsody	✓			
(Sitaram et al., 2016)	SSW			✓	
(Soto and Hirschberg, 2017)	Interspeech	✓			
(Ramanarayanan and Suendermann-Oeft, 2017)	Interspeech	✓			
(Guzmán et al., 2017)	Interspeech	✓			
(Bullock et al., 2018b)	Interspeech	✓			
(Soto et al., 2018)	Interspeech	✓			
(Soto and Hirschberg, 2019)	Interspeech	✓			
(Chandu and Black, 2020)	Interspeech	✓			

Table 20: ISCA Catalog in European-English.

Paper	Proceeding	Modern Standard Arabic
(White et al., 2008)	Interspeech	✓
(Ali et al., 2021)	Interspeech	✓
(Chowdhury et al., 2021)	Interspeech	✓

Table 21: ISCA Catalog in Middle Eastern-English.

Paper	Proceeding	Frisian-Dutch	Russian-Ukrainan
(Lyudovyk and Pylypenko, 2014)	Interspeech		✓
(Yılmaz et al., 2016)	Interspeech	✓	
(Yılmaz et al., 2017b)	Interspeech	✓	
(Yılmaz et al., 2017a)	Interspeech	✓	
(Yılmaz et al., 2018b)	Interspeech	✓	
(Yilmaz et al., 2018c)	SLTU	✓	
(Wang et al., 2019)	Interspeech		✓
(Yılmaz et al., 2019)	Interspeech		✓

Table 25: ISCA Catalog in Two Languages in the same family.

Paper	Proceeding	Kazakh-Russian	Hindi-Tamil	French-Arabic
		1	1	4
(Amazouz et al., 2017)	Interspeech			✓
(Thomas et al., 2018a)	Interspeech		✓	
(Wottawa et al., 2018)	Interspeech			✓
(Chandu and Black, 2020)	Interspeech			✓
(Chowdhury et al., 2021)	Interspeech			✓
(Mussakhojayeva et al., 2022a)	Interspeech	✓		

Table 26: ISCA Catalog in Two Languages in different families.

Paper	Proceeding	SEA Mandarin-English	African Languages-English	Indian Languages-English	Others
		17	1	1	7
(Badino et al., 2004)	Interspeech				✓
(Oria and Vetek, 2004)	Interspeech				✓
(Marcadet et al., 2005)	Interspeech				✓
(Romsdorfer and Pfister, 2006)	ML				✓
(Lyu et al., 2010b)	Interspeech	✓			
(Imseng et al., 2010)	Interspeech				✓
(Weiner et al., 2012b)	SLTU	✓			
(Adel et al., 2014c)	Interspeech	✓			
(Adel et al., 2014b)	Interspeech	✓			
(Giwa and Davel, 2014)	Interspeech		✓		
(Adel et al., 2014a)	SLTU	✓			
(Rallabandi and Black, 2017)	Interspeech			✓	
(Chandu et al., 2017)	Interspeech				✓
(Garg et al., 2018c)	Interspeech	✓			
(Xu et al., 2018)	Interspeech	✓			
(Guo et al., 2018)	Interspeech	✓			
(Chang et al., 2019)	Interspeech	✓			
(Khassanov et al., 2019)	Interspeech	✓			
(Lee et al., 2019b)	Interspeech	✓			
(Zeng et al., 2019)	Interspeech	✓			
(Hu et al., 2020)	Interspeech	✓			
(Li and Vu, 2020)	Interspeech	✓			
(Zhou et al., 2020)	Interspeech	✓			
(Nekvinda and Dušek, 2020)	Interspeech				✓
(Qiu et al., 2020)	Interspeech	✓			
(Liu et al., 2021)	Interspeech	✓			

Table 28: ISCA Catalog in 4+.