

# Arguing Brexit on Twitter

a corpus-linguistic study

Natalie Dykes, Philipp Heinrich  
and Stefan Evert

*Chair of Computational Corpus Linguistics*  
**Friedrich-Alexander-Universität Erlangen-Nürnberg**

June 27, 2019



FRIEDRICH-ALEXANDER  
UNIVERSITÄT  
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT  
UND FACHBEREICH THEOLOGIE

# Argument mining

- relatively new field in NLP and CSS
- goal: automatic extraction and representation of arguments from texts
  - ▶ find argumentative parts of the text
  - ▶ identify premises and conclusions
  - ▶ map text snippets to argumentation schemes
- challenges when looking at real texts:
  - ▶ lots of (defeasible) forms of argumentation (Walton et al., 2008)
  - ▶ premises – or even conclusion – are often left implicit (Bosc et al., 2016)
  - ▶ persuasion through rhetoric strategies such as selection, arrangement, phrasing of argumentative units (Wachsmuth et al., 2018)

# Challenges when looking at social media

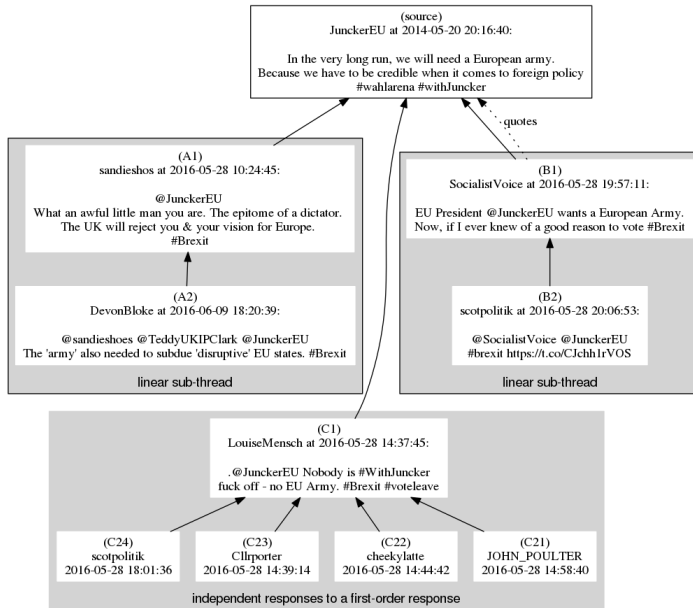
- peculiarities of social media data:
  - ▶ *ad hominem* in various creative flavours
  - ▶ high degree of implicitness (on Twitter: 140 characters)
  - ▶ unmediated environment (no structure or guidelines, multimodality)
  - ▶ non-standard language (CMC)
- automatically finding and classifying arguments in social media is really difficult (Goudas et al., 2014; Dusmanu et al., 2017):
  - ▶ detecting argumentative parts:  $F_1 \approx 80\%$
  - ▶ splitting claims and premises:  $F_1 \approx 40\%$
  - ▶ distinguishing facts from opinions:  $F_1 \approx 80\%$

## Related work

- Degano (2007) explores indicators for presuppositions and dissociations; starts from a predefined list of explicit markers (cf. Levinson, 1983: 181–184):
  - ▶ clause types
  - ▶ comparisons
  - ▶ descriptions
  - ▶ particular verbs
- O'Halloran (2011) analyses argumentation in transcripts of oral conversations
  - ▶ manual coding of utterances as claims and challenges
  - ▶ quantitative part: statistical *keywords*, *key POS tags* and *key semantic domains* to explore the linguistic realisation of arguments (cf. Rayson, 2008)
- more content-centred approaches based on keyness: exploration of argumentation as one of various contexts (Partington, 2003; Baker, 2004; Al-Hejin, 2015)

# Corpus

- basis: tweets containing “Brexit” (16/05/05–16/08/24) + reply threads
  - ▶  $\approx$  6 million tweets
- pre-referendum originals (higher consistency in argumentation)
  - ▶  $\approx$  2.5 million tweets
- deduplication (Schäfer et al., 2017)
  - ▶  $\approx$  1.8 million tweets, 32 million tokens
- linguistic annotation (several pipelines):
  - ▶ tokenization
  - ▶ part-of-speech tagging
  - ▶ lemmatization
  - ▶ phrase-chunking and NER
- indexing with the IMS Open Corpus Workbench (CWB)



## Some examples

- @DPJHodges @NickCohen4 I'll be voting Brexit because it's what the establishment AND the left don't want.
- @Rachael\_Swindon @mcpeake\_liam and Dyson's for BREXIT because he'd like to replace British & EU engineers with cheaper Chinese ones
- Then again people say they will vote but when it comes to the day they are too lazy to vote #Brexit #VoteLeave
- So if donald trump is backing the brexit campaign id say thats good sign we should all be backing the stay campaign
- @PauleastHull @iVoteStay whatever brexit voters are stupid, racist, brainwashed and gullible, which of those 4 are you
- @JMoneyMC brexit is really just the gateway towards european nationalism so yea lol, it's fucking racist

# Patterns and schemes

$\backslash\text{forall } \{?0 : \text{entity}\} \backslash\text{in } \{?1 : \text{entity}\} : \{?2 : \text{property}\}(?0)$

$\iff$  “all entities in entity 1 have property 2, e.g. being an idiot”

- Common Folks *ad populum* (Walton et al., 2008: 129):
  - ▶ premise: I (the speaker) am an ordinary person, that is, I share a common background with you (the audience).
  - ▶ conclusion: Therefore, you ought to take what I say as being more credible or acceptable.
- Position to Know *ad populum* (Walton et al., 2008: 126):
  - ▶ premise 1: Everybody in group  $G$  accepts  $A$ .
  - ▶ premise 2: Group  $G$  is in a special position to know that  $A$  is true.
  - ▶ Conclusion: Therefore,  $A$  is (plausibly) true.



# Schemes and instances

- Common Folks *ad populum*:
  - ▶ @paulmasonnews @Medium Amen brother , voting to stay , because **the average person doesn't** need to be left in the hands of the brexit leaders !!
  - ▶ @DrAlanGreene I'm as against #Brexit as **the next man** but this is nonsense . Why go via Ireland rather than Calais ? Ppl reach Calais by train .
- Position to Know *ad populum*:
  - ▶ @vincentemery95 You're wrong , yet again . Any excuse to be anti European . Fortunately most **business people like myself** know better .
  - ▶ @OffencePolice @MarcherLord1 @OHStillHere **as an expat** ( and a Scot .... ) i'm pulling for #Brexit . Screw the Remainiacs ...

# Corpus linguistics for ArgMining

- schemes have different linguistic representations
- linguistic patterns, but not directly tied to word level
- corpus-linguistic approach via CQP query language (Evert and Hardie, 2011)
  - ▶ specification of phrase chunks
  - ▶ word lists representing lexico-grammatical categories
  - ▶ iterative development informed by regular analysis of concordance lines, e. g.:
    - ★ #EURef @connor\_p\_obrien I have read the news recently and **as a result I have mourned** . What should I be scared of , then ?
  - ▶ goal: high precision (low recall)

# Corpus queries

as an X I Y

```
[lemma = "as"] <np>[*
@0[
    lemma = $person_common | lemma = $person_profession |
    lemma = $person_origin | lemma = $person_support
] [* </np> [lemma = "i"] <vp>[*
@1[pos_simple="V"]
@2[*</vp> (
    <vp> [* </vp> | <np>[* [pos_simple="N|Z"] [*</np> |
    /be_ap[* | <pp>[* [pos_simple = "N|Z"] [*</pp> |
    [pos_simple="R"] | [word != "."])*
)+
@3::] within tweet;
```

# Word lists: nouns\_person\_negative

[a-z]*ass	cheater	elite	gutter	lot	prick	stunt
asshole	child	expert	guy	maniac	puppet	swine
bandit	clown	extremist	hater	[a-z]*migrant	pussy	thug
bastard	contempt	fag	hipster	millenial	pygmy	toad
bat	crap	faggot	hysteric	monkey	racist	toddler
bean	cretin	fart	idiot	moron	rapist	tory
bitch	criminal	fascist	incompetent	muppet	rat	traitor
bigot	crowd	fathead	infiltrator	murderer	redneck	tranny
bint	cult	folk	joke	neanderthal	remainer	trash
boy	cunt	fool	kid	nigger	#?remaniac	troll
brexiter	denier	[a-z]*fucker	knobhead	numpty	retard	twat
brigade	dick[a-z]*	fucknut	leaver	nut	scaremongering	voter
bugger	dictator	fuckup	leftist	nutter	scoundrel	wank
bully	dinosaur	gambler	lefty	peasant	scum	wanker
bunch	disdain	garbage	liar	[a-z]+phile	scumbag	weirdo
butthole	disgrace	gimp	lib	piece	sheep	wreck
bureaucrat	dodo	girl	liberal	pig	[a-z]*shit	whore
camp	douche	globalist	load	politician	shite	xenophobe
campaigner	douchebag	gobshite	loon	populist	sod	yank
caveman	dummy	goof[a-z]*	loser	pretender	soul	zombie

## Case Study: Self-identification in the Brexit discourse

$\forall x \{x : \text{entity}\} \in \{y : \text{entity}\} : \{z : \text{property}\}(x)$

$\iff$  “all entities in entity 1 have property 2, e.g. being an idiot”

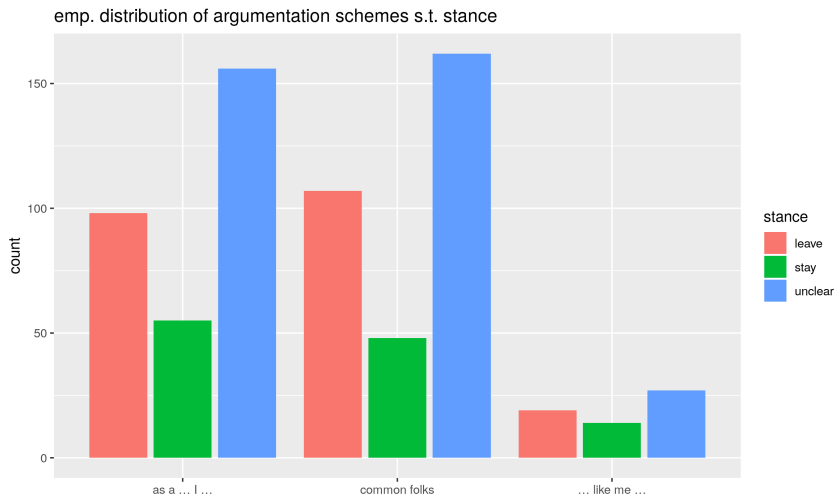
- three query types:

- ▶ as an  $X$  I  $Y$
- ▶  $X$  like me  $Y$
- ▶ {ordinary/normal/common} {people}  $Y$

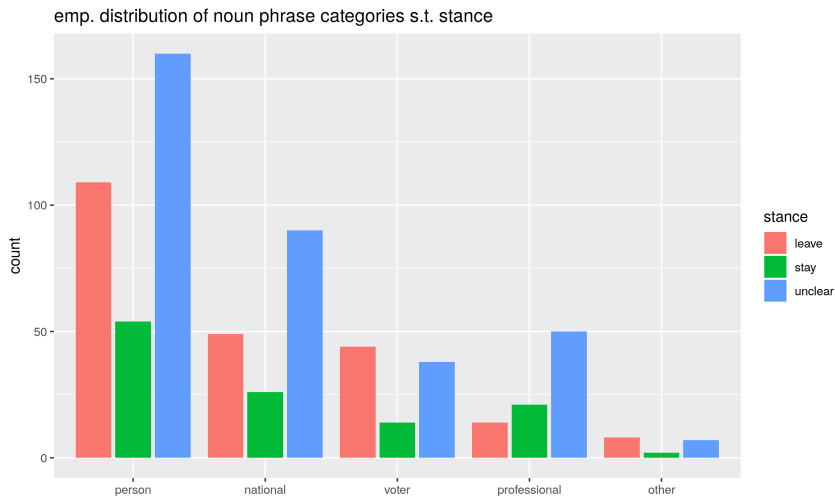
- manual categorization:

- ▶ leave – stay – unclear
- ▶ group identity statements categorised by domains
  - ★ majority: no clear stance
  - ★ stance may differ within group membership

# Case Study: Self-identification in the Brexit discourse



# Case Study: Self-identification in the Brexit discourse

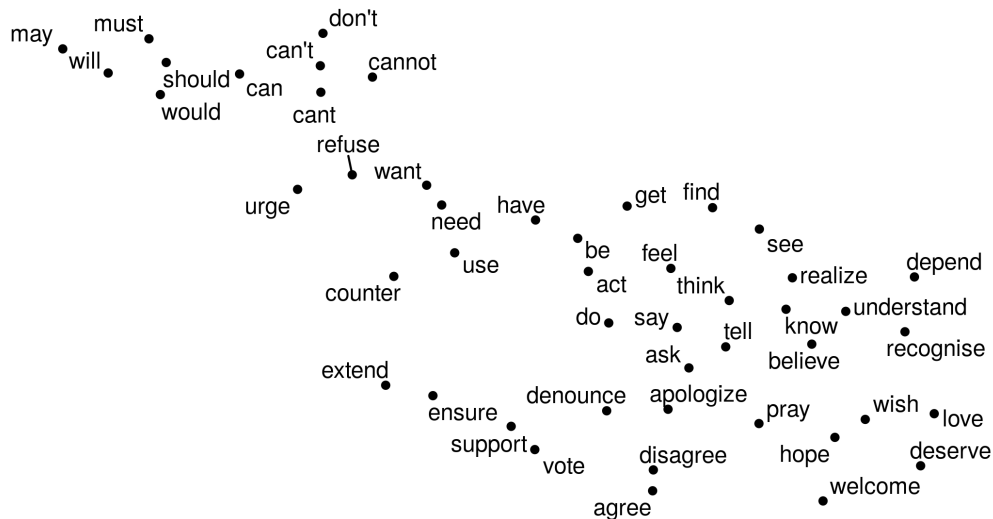


# Query: as an $X \mid Y$

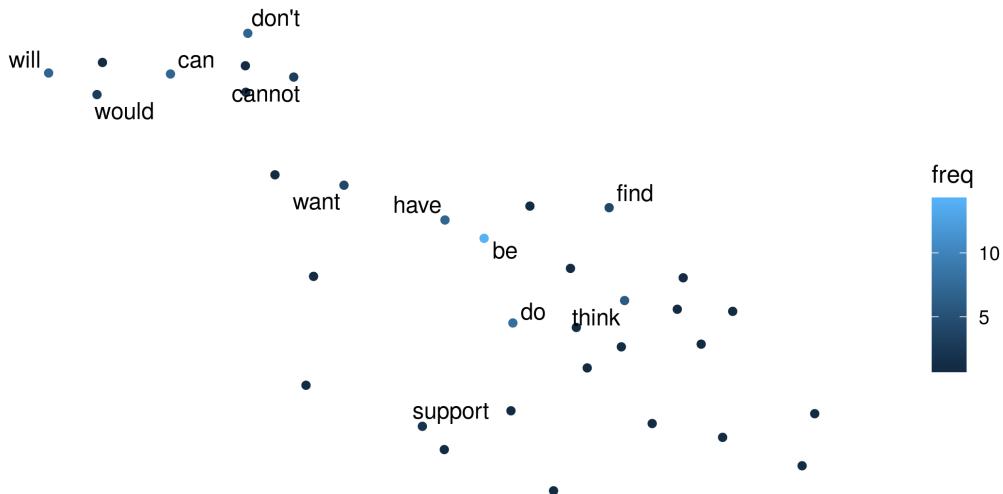
```
[lemma = "as"] <np>[]*
@0[
  lemma = $person_common | lemma = $person_profession |
  lemma = $person_origin | lemma = $person_support
] []* </np> [lemma = "i"] <vp>[]*
@1[pos_simple="V"]
@2[]*</vp> (
  <vp> []* </vp> | <np>[]* [pos_simple="N|Z"] []*</np> |
  /be_ap[] | <pp>[]* [pos_simple="N|Z"] []*</pp> |
  [pos_simple="R"] | [word != "."]*
)+
@3[:,:] within tweet;
```



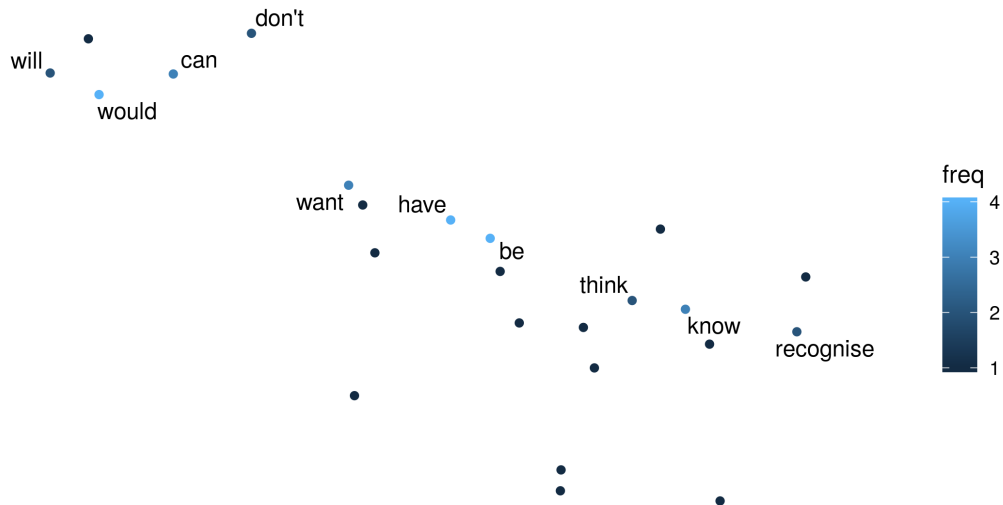
query: 'as\_x\_i\_y', verbs in the Y-slot



group: support (94 instances)



group: profession (43 instances)



group: common (35 instances)



## Some concordance lines

- EU is a social experiment on the redistribution of wealth **as a socialist I support** the theory but its not working so I'm voting out #Brexit
- @JohnDowson3 @nickeardleybbc @theSNP She 's not wrong regarding the fishing , **as a former fisherman i know** the UK government failed us .
- @StrongerInPress That 's why **as a young person I am voting to leave** you utter cockwombles . #Brexit #VoteLeave #Votin

# Conclusion and Future Work

## Conclusion

- queries balance grammatical and semantic flexibility in patterns
- each query: one linguistic instantiation of a given argumentation scheme
- qualitatively informed approach to handle noisy data

## Future work

- interactive toolkit
- mapping of schemes to logical formulae (theoretical CS)

Presenters like us would like to thank you for listening.

As FAU researchers we are looking forward to your questions.



- B. Al-Hejin. Covering Muslim women: Semantic macrostructures in BBC News. *Discourse & Communication*, 9(1):19–46, 2015.
- P. Baker. 'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics*, 8(1):88–106, 2004.
- Tom Bosc, Elena Cabrio, and Serena Villata. Tweeties squabbling: Positive and negative results in applying argument mining on social media. In *Proceedings of the 6th International Conference on Computational Models of Argument*, 2016.
- C. Degano. Presupposition and Dissociation in Discourse: A corpus study. *Argumentation*, 21: 361–378, 2007.
- Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- Stefan Evert and Andrew Hardie. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*, 2011.
- Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham, 2014. Springer.



- S. Levinson. *Pragmatics*. Cambridge textbooks in linguistics, Vol. 8. Cambridge University Press, 1983.
- K. O'Halloran. Investigating argumentation in reading groups: Combining manual qualitative coding and automated corpus analysis tools. *Applied Linguistics*, 32(1):172–196, 2011.
- A. Partington. *The linguistics of political argument: The spin-doctor and the wolf-pack at the White House*. Routledge, 2003.
- P. Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008.
- Fabian Schäfer, Stefan Evert, and Philipp Heinrich. Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism and PM Abe Shinzō's Hidden Nationalist Agenda. *Big Data*, 5(4):1 – 16, 2017.
- Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, 2018.
- D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.