Annotating speaker stance in discourse: the Brexit Blog Corpus (BBC)

SND-ID: snd1037-1. **Version**: 1.0. **DOI**: https://doi.org/10.5878/002925

Ladda ner data

brexit_blog_corpus_raw.zip *(6.76 MB)*

brexit_blog_corpus.xlsx *(151.42 KB)*

Ladda ner alla filer

snd1037-1-1.0.zip *(~6.91 MB)*

Citering

Kerren, A., & Paradis, C. (2017) Annotating speaker stance in discourse: the Brexit Blog Corpus (BBC) (Version 1.0) [Dataset]. Linnéuniversitetet. Tillgänglig via: https://doi.org/10.5878/002925

Skapare/primärforskare

Andreas Kerren - Linnéuniversitetet

Carita Paradis - Lunds universitet, Språk- och litteraturcentrum

Forskningshuvudman

Linnéuniversitetet - Institutionen för datavetenskap

Beskrivning

I studien har man undersökt i vilken utsträckning språkanvändare är överens om vilka ståndpunkter som uttrycks i vardagligt språk eller om tolkningarna skiljer sig åt. För att utföra denna uppgift utvecklades ett omfattande kognitivt-funktionellt ramverk bestående av tio kategorier som representerade olika inställningar och som baserades på tidigare arbeten om talares uppfattning som finns i litteraturen. En korpus av åsiktsladdade texter, där talare tar ställning och positionerar sig, sammanställdes genom The Brexit Blog Corpus (BBC). Ett analytiskt gränssnitt för annoteringarna upprättades och data annoterades av två oberoende annotatorer. Annoteringsförfarandet, överenskommelsen om hur annoteringen skulle bedrivas och förekomsten av mer än en inställningskategori bland de studerade uttalandena finns beskrivna. Den noggranna analytiska annoteringsprocessen har hög utsträckning lett till tillfredsställande inter- och intra-annoteringar, vilket i den slutliga versionen av BBC resulterade i en guldstandardkorpus

Syfte:

Syftet med studien är att undersöka om det är möjligt att identifiera olika talares inställning i diskursen genom att tillhandahålla en analytisk resurs för detta och därefter utvärdera nivån av enighet mellan olika talare i diskursen.

BBC är en samling av texter som hämtats från bloggar. Korpustexterna är tematiskt relaterade till den brittiska folkomröstningen 2016 som gällde huruvida Storbritannien borde förbli medlemmar i Europeiska unionen eller ej. Texterna extraherades från Internet under perioden juni till augusti 2015. Med Gavagai API (https://developer.gavagai.se) hittades texterna med hjälp av nyckelord som: Brexit, EU referendum, pro-Europe, europhiles, eurosceptics, United States of Europe, David Cameron, eller

Downing Street. URL:erna som hämtades filtrerades så att endast engelska sidor som beskrivs som bloggar valdes. Varje nedladdad dokument delades upp i sententiella uttalanden, varav 2 200 uttalanden valdes slumpmässigt för analysen. Den slutliga storleken på korpusen är 1 682 uttalanden, 35 492 ord (169 762 tecken utan mellanslag). Varje uttalande innehåller mellan 3 och 40 ord med en medellängd på 21 ord.

För dataannoteringsförloppet användes verktyget the Active Learning and Visual Analytics (ALVA) (https://doi.org/10.1145/3132169 och https://doi.org/10.2312/eurp.20161139). Två annotatorer, varav den ena är en professionell översättare med licentiatexamen i engelsk lingvistik och den andra har en doktorsexamen i beräkningslingvistik, utförde annoteringarna oberoende av varandra.

Datasetet kan laddas ned i två olika format: antingen som Excel-fil eller i ett rådatabasformat (ZIP-arkiv) som kan vara användbart för analytiska ändamål och maskininlärning, till exempel med Python-biblioteket scikit-learn. Excel-filen innehåller ytterligare en variabel (utterance word length). ZIP-arkivet innehåller en uppsättning kataloger (t.ex. "contrariety" och "prediction") som motsvarar inställningskategorierna. Inuti varje sådan katalog finns två kataloger som motsvarar annoteringar som tilldelar eller inte tilldelar respektive kategori som uttalanden (t.ex. inom den överliggande kategorin "prediction" finns det två underliggande kataloger, där den ena heter "prediction" och innehåller uttalanden som märkts med denna kategori, och "no" som innehåller resterande uttalanden). Inne i katalogerna finns det textfiler som innehåller individuella uttalanden.

Vid användande av data från den här studien önskar primärforskaren att citering också görs till publikationen: Vasiliki Simaki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher, and Andreas Kerren. Annotating speaker stance in discourse: the Brexit Blog Corpus. In Corpus Linguistics and Linguistic Theory, 2017. De Gruyter, published electronically before print. https://doi.org/10.1515/cllt-2016-0060

Språk

Engelska

Analysenhet

Mediaenhet: Text

Tidsperiod(er) som undersökts

2015-06-01 – 2016-05-31

Variabler

8

Antal individer/objekt

1682

Dataformat / datastruktur

Text

Datainsamling 1

- Tidsperiod(er) för datainsamling: 2015-06-01 – 2016-05-31
- Datakälla: Forskningsdata

Ansvarig institution/enhet

Institutionen för datavetenskap

Finansiering

- Finansiär: Vetenskapsrådet
- Diarienummer hos finansiär: 2012-5659

Forskningsområde

Informationsteknik (CESSDA Topic Classification)

Språkteknologi (språkvetenskaplig databehandling) (Standard för svensk indelning av forskningsämnen 2011)

Jämförande språkvetenskap och allmän lingvistik (Standard för svensk indelning av forskningsämnen 2011)

Studier av enskilda språk (Standard för svensk indelning av forskningsämnen 2011)

Språk och lingvistik (CESSDA Topic Classification)

Media, kommunikation och språk (CESSDA Topic Classification)

Nyckelord

Textannotering, Blogtexter, Modalitet, Opinion, Känsloyttring, Värdering, Bedömning, Evidentialitet, Subjektivitet, Attityd, Positionering

Publikationer

Vasiliki Simaki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher, and Andreas Kerren. Annotating speaker stance in discourse: the Brexit Blog Corpus. In Corpus Linguistics and Linguistic Theory, 2017. De Gruyter, published electronically before print. https://doi.org/10.1515/cllt-2016-0060

Om du publicerat något baserat på det här datamaterialet, meddela gärna SND en referens till din(a) publikation(er). Är du ansvarig för katalogposten kan du själv uppdatera metadata/databeskrivningen via DORIS.

Tillgänglighetsnivå

Åtkomst till data via SND
Data är fritt tillgängliga

Användning av data

Att tänka på vid användning av data som delas via SND

Versioner

Version 1.0. 2017-10-13

Hemsida

Hemsida

Kontakt för frågor om data

Andreas Kerren

andreas.kerren@lnu.se

CLARIN Virtual Collection Registry

Lägg till i samling

En virtuell samling är kopplad till ett specifikt forskningsändamål och innehåller länkar till dataresurser i olika digitala arkiv. Samlingen är lätt att skapa, få åtkomst till och citera.

Read more about virtual collections on the CLARIN website.

Ladda ner metadata

DataCite

DDI 2.5

DDI 3.3

DCAT-AP-SE 2.0

JSON-LD

PDF

Citering (CLS)

Filöversikt (CSV)

Publicerad: 2017-10-13
Senast uppdaterad: 2019-01-15