Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

Samenvatting

Computerondersteunde corpuslinguïstiek is een van de belangrijkste punten van convergentie tussen taalkundige en computationele methoden. In het bijzonder biedt het gebruik van diachronische linguïstische corpora kansen voor de kwantitatieve en computationele analyses van fenomenen met betrekking tot taalverandering door de tijd heen. Dit proefschrift presenteert het resultaat van drie onafhankelijke, hoewel gerelateerde, projecten die dezelfde interesse hebben in de toepassing van rekenkracht op diachronische corpuslinguïstiek. Het belangrijkste doel is bijdragen te leveren aan drie van de fasen van het onderzoek waarbij diachronische taalkundige corpora betrokken zijn: (a) opbouw en compilatie van corpora; (b) ontwerp van instrumenten en algoritmen voor data-exploratie; en (c) gegevensanalyse voor taalkundig, cultureel en historisch onderzoek. Twee nuttige bronnen voor corpuslinguïstiek in een andere taal dan het Engels worden eerst gepresenteerd: een webschraper van commentaren van nieuwsportalen en websites, ontwikkeld als open source en als vrij te gebruiken, te wijzigen en te verspreiden; en een vrij verkrijgbaar diachronisch corpus samengesteld uit commentaren gepubliceerd op UOL, een belangrijk Braziliaans nieuwsportaal. Deze bronnen zijn niet alleen relevant voor taalkundigen, maar ook voor professionals uit andere

vakgebieden, waaronder sociale wetenschappers en journalisten die zich bezighouden met de publieke perceptie van nieuws en de relatie tussen de media en de samenleving. Vervolgens stellen we een eenvoudige en generaliseerbare methode voor om de periodes van oprichting en veroudering van linguïstische items in een diachronisch corpus te helpen identificeren op basis van de frequentie van deze items in het corpus. Deze methode kan worden gebruikt voor de analyse van elke verzameling van taalkundige items, ongeacht de taal of de historische periode. We tonen ook de toepasbaarheid aan met behulp van lexicale gegevens van het Corpus of Historical American English (COHA), met casestudies over de statistieken en kenmerken van woorden die in verschillende periodes in dit corpus voorkomen of uit dit corpus verdwijnen. Tot slot beschrijven we hoe diachronische corpora kunnen worden gebruikt voor kwantitatief taalkundig onderzoek door in eerste instantie een kader voor te stellen dat gericht is op het onderzoek van de woordenschat door middel van een diachronische benadering die gebruik maakt van complementaire methoden uit de corpuslinguïstiek en natuurlijke taalverwerking. De toepasbaarheid van dit kader wordt vervolgens aangetoond door de casusanalyse van de term *fake news*, waarvan we de perceptie en conceptualisering in de traditionele media onderzoeken met behulp van gegevens verzameld uit Braziliaanse en Engelstalige mediabronnen. Met deze bijdragen verwachten we onderzoek naar diachronische corpuslinguïstiek en naar computationele methoden voor taalkundige analyse te bevorderen.

Trefwoorden: corpus linguïstiek; computerondersteunde linguïstiek; diachronische studies.