The new 4.3 billion word NOW corpus, with 4--5 million words of data added every day

Mark Davies (Brigham Young University, USA)

One of the challenges facing corpus creators and users is the fact that so many corpora quickly become "stale". They may do a great job of representing the language from 10--20 years ago, but there is nothing from the last year or two – or especially the last month and definitely not from yesterday (see Davies 2011).

In May 2016 we released the NOW corpus ("Newspapers on the Web), which is meant to help fill this gap (http://corpus.byu.edu/now/). The corpus is composed of more than 4.3 billion words of data from newspapers from twenty different countries, from January 2010 to the current time. (Because the corpus is continually growing, it will be more than 4.7 billion words in size by the time the CL 2017 conference is held in July 2017.)

Most importantly, however, automated scripts run every day to add texts to the corpus. Each day these scripts 1) get 10,000--15,000 URLs from Google News 2) download the web pages with HTTrack 3) clean them up with JusText to remove boilerplate material 4) tag and lemmatize the texts with CLAWS 7, 5) remove duplicates (based on n--grams), and then 6) integrate the texts into the existing relational database architecture.

In this way, there are 4--5 million words from approximately 9,000--10,000 new texts every day. For example, this abstract is being written the morning of April 20, 2017 and the corpus already contains 5.5 million words (in 11,700 articles) from yesterday – April 19, 2017. This daily expansion of the corpus translates into about 130 million words each month, and about 1.5 billion words of data each year.

With the NOW corpus, users can see what is happening in the language this week ---- not just 10 or 20 years ago. For example, they can find the most recent 100 hits for any word or phrase, and in many cases they will have hits from yesterday. Such data is typically much more relevant and interesting to language learners than data from the early 1990s, before they were even born.

The following are just a handful of words that probably would only be in a corpus that includes very recent texts, and all of which have many tokens in NOW: (general words) Brexit (n) 2015, manspreading (n) 2015, makerspace (n) 2015, gig economy (n) 2015, dadbod (n) 2015, momager (n) 2015, swatting (n) 2015, walkscore (n) 2015, trigger warning (n) 2014, mommy porn (n) 2014, normcore (n) 2014, listicle (n) 2014, sufferfest (n) 2014, and catfishing (n) 2013;; (computers / technology / science) Uberization (n) 2015, selfie stick (n) 2015, fracklog (n) 2015, digital detox (n) 2015, droneport (n) 2015, data lake (n) 2015, smartwatch (n) 2014, and airpocalypse (n) 2013. Hundreds of examples of similar recent words can be found at the NOW website.

Of course the corpus allows users to do more than just search for individual words and phrases. They can find families of new words, such as words containing *fest (more than 2,600 different types), e.g. horrorfest, borefest, smashfest;; *sexual* (more than 1,600 types), e.g. metrosexual, demisexual, pansexual;; *phobia (more than 900 types), e.g. photophobia, dronophobia, robophobia;; smart* (more than 2,800 types), e.g. smartscooter, smartwater, smartart;; *ware

(more than 600 types), e.g. adware, crapware, Slackware;; *athon (more than 780 types), e.g. laughathon, funathon, stavathon;; *geddon (more than 230 types), e.g. carmageddon, stormageddon, hairmageddon. and *alypse (nearly 400 types), e.g. apocalyse, snowpocalypse, zombocalpyse. Data such as this allows us to look at morphological productivity and lexical creativity in real time, since the corpus is never more than 24 hours out of date.

Because the BYU corpus architecture allows users to easily and quickly compare between different sections of the corpora, in just 2--3 seconds a researcher could find, for example, words occurring with digital NOUN that are much more common in 2015--2017 than in 2010--2012, e.g. digital nodes, digital insights, digital mall, digital factory, digital maturity, and digital chaos. A similar search for data NOUN in 2015--2017 vs 2010--2012 would yield data front, data sale, data leak, data fabric, data ingestion, or data dude and hundreds of other new phrases.

In addition to seeing the frequency in six month segments, users can also see the frequency of words and phrases by week, to see when a particular topic was discussed the most since 2010. They can also find the keywords for a given day (including yesterday), week, month, or year. For example, they could find the keywords for Apr 4 2016 (related to the Panama papers leaks), Mar 22 2016 (the Brussels bombing and Obama going to Cuba), and Nov 13 2015 (the Paris attacks).

With updates to the architecture and interface of the BYU corpus interface in May 2016, it is now possible to quickly and easily create "virtual corpora" within a given corpus (like NOW). For example, in just 5--10 seconds users could create virtual corpora based on the following words: investment*, rebound (i.e. basketball), or electron. They can also create virtual corpora based on date and source (e.g. New York Times in Sep-Dec 2016).

For example, in just 10--20 seconds, a user could create a 600,000--700,000 word corpus based on texts from the UK in September 2015 dealing with refugees in Europe. The user could then search within the virtual corpus to find any word or phrase, such as asylum, or to find the collocates of a given word. They could also compare the frequency of a word or phrase in multiple virtual corpora, such as "refugee" corpora from the US, UK, Canada, and Australia in the same time period. Finally, in just 5--10 seconds they can generate "keyword" lists from the virtual corpora, e.g. refugee, migrant, asylum, seeker, influx, border, quota, crisis, fence, camp, arrival, boat, and shelter. Obviously, this allows researchers to quickly and easily obtain data on very contemporary topics.

In summary, a corpus like NOW allows us to move far beyond moderately--sized 10--20 year old corpora, to examine billions of words of data and see language change as it occurs.

References

Davies, M. (2011). The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, 25(1), 447-65.