Corpus Linguistics and Digital Humanities. Intersecting Paths. A Case Study from Twitter

América Crítica

América Crítica 4(2), 131–141, 2020, ISSN: 2532-6724, https://doi.org/10.13125/americacritica/4521

Corpus Linguistics and Digital Humanities. Intersecting Paths. A Case Study from Twitter

Angela Zottola[1]

[1] Università di Torino, Italy

Received: 17/08/2020

Accepted: 10/11/2020

Abstract—In this paper I aim at critically discussing the role of Corpus Linguistics within the field of Digital Humanities. I posit that the accessible and user-friendly tools of Corpus Linguistics are an optimal resource for scholars within the Humanities and the Social Sciences to engage in Digital Humanities and take in its effort to bring computing techniques to humanities research even further. I also present a case study based on data collected from Twitter as an example of how the two approaches can come together within the framework of American Studies. In this paper American Studies is conceived as a discipline inclusive of any perspective that looks at the American continent rather than a specific field of research. I conclude by endorsing the crossing paths between Digital Humanities and Corpus Linguistics as a necessity in the future of Digital Humanities. — *Digital Humanities, corpus linguistics, American Studies, Twitter, WMatrix*.

Abstract—L'obiettivo di questo saggio è proporre un'analisi critica del ruolo della Linguistica dei Corpora all'interno delle Digital Humanities. Lo studio suggerisce che gli strumenti accessibili e user-friendly della Linguistica dei Corpora possono essere considerati uno strumento ottimale attraverso il quale operare all'interno delle Scienze Umane e avanzare nell'obiettivo delle Digital Humanities di includere varie tecniche di computazione negli studi umanistici. Il saggio presenta un caso studio basato su dati raccolti sul social medium Twitter che rappresenta un esempio in cui i due approcci (Linguistica dei Corpora e Digital Humanities) possono essere utilizzati unitamente all'interno degli American Studies. Nel saggio, gli American Studies sono intesi in senso lato come una disciplina inclusiva di diverse prospettive che si approcciano allo studio del continente americano. In conclusione, l'incontro tra la Linguistica dei Corpora e le Digital Humanities viene definito come una necessità per il futuro delle Digital Humanities stesse. — *Digital Humanities, linguistica dei corpora, American Studies, Twitter, WMatrix*.

Introduction

In this paper I aim to discuss the existing and possible developments in the intersection between Digital Humanities (DH) and Corpus Linguistics (CL), arguing that the type of tools offered by CL, which combine quantitative and qualitative analyses, can be a valuable asset to a number of disciplines within the Humanities and Social Sciences – disciplines inherently located within the qualitative spectrum - as a way to approach the DH. CL and DH are both technology-mediated approaches widely used in both the Humanities and Social Sciences to investigate how language produced in different settings is employed to construct meaning. However, it could be argued that while CL is a linguistic-based framework that uses technology as a way to assist scholars in the analysis of language and is mainly used by researchers interested in researching how language

Contact data: Angela Zottola, angela.zottola@unito.it

works by investigating its structures, patterns and the intrinsic characteristics of its use, scholars who position themselves within DH are not necessarily language experts. In DH technology has a more central role and the study of language is often auxiliary to the investigation of digital data.

In this study, I argue that while DH and CL are well-established frameworks of their own, and have many overlapping goals and instruments, the combined use of the two is not very popular, especially in research which is not strictly related to the field of linguistics. It is rather rare to find a paper using CL as its methodological framework and outwardly positioned within the field of DH. At the same time, most research framed at the intersection between DH and linguistics usually applies methods from Computational Linguistics or Natural Language Processing (NLP) (see, among others, McGillivray *et al*. 2020; Arnold *et al*. 2019; Sprugnoli *et al*. 2019).

My work stems from a background in linguistics and in this paper I will suggest that not only the field of linguistics could benefit from a stronger and more explicit connection to DH, but most importantly that the combination of CL techniques within the DH framework could be an asset for other fields within the Social Sciences. While this discussion has been started in the past (see Jensen 2014) and increasingly continued in the past few years (Brookes and McEnery 2020), I would also like to offer an additional layer to this discussion by including within the conversation the field of American Studies.

In fact, this paper is set out to provide a definition of the two frameworks and discuss the increasing importance of bringing them together as a way to make DH more accessible to scholars less familiar with digital practices. It also presents a small case study based on data collected through Twitter and that, given its geographical, political and social implications, can be positioned within the framework of American Studies, conceding that this is considered from a broad theoretical perspective. As Brinson Curiel *et al*. (2000: 14) have suggested, one of American Studies' traditional goal is "interdisciplinary thinking about American experience." In fact, the field has a longstanding and established cross-disciplinary practice. Following up on this practice, American Studies offers a suitable ground for scholars interested in America[1] to go beyond those areas in which it has already flourished such as history and literary criticism (Brinson Curiel *et al*. 2000, 14) and be more inclusive, both in terms of objects of studies but also of the approaches and disciplines that can be included under the American Studies label. This is especially true in Italy, for at least three reasons. First of all, the status of American Studies is unfortunately ambiguous due to its absence as an "institutional field in the Italian university system" (Izzo 2018: 184). Secondly, not many linguists, despite their interest in the United States, commonly relate to the field - to the best of my knowledge I have yet to find a paper which positions itself at the intersection between CL (or other linguistic-based approaches) and American Studies. And lastly, the DH are just recently making way in universities around the country. In this sense, I want to reproduce here what has been appropriately said by Simon Bronner (2012), to summarize this paper's understanding of the field. Bronner argues that the "matrix" principle for American Studies is that it can be seen as

> a location for progressive research, a form of area studies, allowing in its flexible domain multiple ways of viewing the same subject- the Unites States or the Americas – and forging integrated approaches that could be called inter- or transdisciplinary. An alternative view is that American Studies is counterdisciplinary because it is problem centered in a reform project of the bureaucratic university and works to break down departmental walls. (n.p.)

With the aim of bringing together American Studies, DH and CL, interdisciplinarity is the *fil rouge* in this paper. Through an interdisciplinary approach, knowledge and the critical perspectives offered by the three fields can intersect and provide novel scholarship, novel perspectives and a new awareness on the way we can explore issues related to society, politics and the importance of language use on a daily basis. By crossing disciplinary boundaries scholars can look at data with different sets of interpretative tools, and, in the case of DH prospects are broadened in terms of the data that is possible to retrieve and collect. This paper is a testament to the richness that is brought by this use of interdisciplinarity, whereas a linguistic analysis is enhanced and supported by theories drawn from Social and Cultural Studies.

In the next sections, I will first introduce the fields of DH and CL, critically addressing their commonalities

[1] Or the Americas as it has been suggested by a number or scholars (see among others Levander and Levine 2007). Due to space constraints this discussion cannot be explored further here but it is without doubt a relevant one considering the idea of a broader and more inclusive field that I refer to.

and differences. I will then introduce a case study carried out on a corpus of Tweets. I will conclude this paper by going back to DH and CL and addressing the reasons that lead me to argue in favor of a growing intersection of the two disciplines.

A Focus on CL and DH

In this section, as a way of introducing the two frameworks, I will discuss CL and DH. As anticipated in the introduction, I view these as two approaches that are at the same time different but similar in some aspect and that are both effective tools for the analysis of language in a digital format. Alexander Dunst (2016) suggests that many scholars within the humanities already engage in digital scholarship daily, whether they identify as digital scholars or not.

As of today, the discussion on a definition of both CL and DH that can be considered as final and upon which everyone agrees is still open among scholar of the two fields. In my view, CL can be considered as a methodological framework, although – as I will discuss further later in this section) – the field is still divided over the choice between theoretical and methodological (Gries 2009), while, despite the longstanding debate on its definition (to read more about this see, among others, Gold and Klein 2016), I identify DH as being closer to a theoretical approach to academic enquiry. The reason behind my characterization of DH as such is related to the fact that the variety of works that identify under this label show that there is no *one* way of doing DH, but there is a vision behind it that guides scholars in this field, in the words of Lisa Spiro (2012) "the digital humanities [seek] to push the humanities into new territory by promoting collaboration, openness, and experimentation" (Spiro 2012: n.p.). What I argue in this paper is that CL should be endorsed as one of the many ways of doing DH, both in linguistics, the field within which CL emerged, but also in other fields which likewise focus on the study of language. In recent times, this suggestion has been increasingly discussed, yet seems not be a given (Brookes and McEnery 2020). In a chapter presenting the Literateca project, Diana Santos (2019) sets out as one of the aims of the text to "test the use of resources and techniques from two different research communities: corpus linguistics and literary digital humanities, complementarily instead of alternatively" (103). In this quote, and in my view, key is the use of the adverb "complementarily." I find CL techniques as a valuable tool for DH for scholars who have different degrees of specialization, as they allow both for more complex statistic-related exploration but also for more basic descriptive observation of the use of language, providing even those with very little familiarity with digital tools with a framework to examine digital data.

Within the field of Linguistics, CL has been at the forefront in the digitalization of research. As suggested by Kim Jensen (2014: 116), "CL has been around for decades and has made leaping and creeping advances in tandem with the development of digital technology." CL has been defined in time in a number of ways, and as Charlotte Taylor posits these include "a *tool*, a *method*, a *methodology*, a *methodological approach*, a *discipline*, a *theory*, a *theoretical approach*, a *paradigm* (theoretical or methodological), or a combination of these." (Taylor 2008: 180) In this work, as suggested earlier, I consider CL as methodological approach useful in the collection and analysis of digital data. We find mentions of CL as early as 1982 (Aarts and van Heuvel 1982; Aarts and Meijs 1984), but the approach became popular only in the early 1990s with scholars such as Leech, Sinclair and Stubbs (Taylor 2008: 179–180). McEnery and Hardie (2012) define CL as "not monolithic, consensually agreed set of methods and procedures for the exploration of language" (2012: 1) mainly "based on examples of real language use." (McEnery and Wilson 1996: 1) All in all, CL can be defined as a heterogeneous and versatile field of inquiry which encompasses a variety of methods and procedures of analysis such as collocation analysis, concordance analysis and keyword analysis, and can be applied in any field that uses natural occurring language as their main source of data.

Corpus, or corpora in its plural declination, is a Latin word that stands for 'body' and it is used in linguistics with reference to a body or collection of written or oral texts. Linguistics has borrowed this term and in modern linguistics it is explained as a collection "of naturally occurring language." (McEnery, Xiao and Tono 2006: 4) According to Gaëtanelle Gilquin and Stefan Gries, corpora, within CL is a collection of texts with very specific features:

- is machine-readable;
- is representative with regard to a particular variety/register/genre, meaning that the corpus contains data for each part of the variety/register/genre the corpus is supposed to represent;
- is balanced with regard to a particular variety/register/genre, meaning that the corpus parts' sizes are proportional to the parts of the vari-

ety/register/genre the corpus is supposed to represent (given the absence of reliable estimates of how much of a target language consists of any one particular variety/register/genre, balancedness is a theoretical ideal);

- has been produced in a natural communicative setting. (2009: 6).

The use of corpus methodologies for language analysis has been considered mostly quantitative, although the numerous relatively recent approaches that combine it with discourse analysis and other methods are changing this orientation towards a mixed quantitative/qualitative methodology. Different types of corpora have been produced in the past twenty years, from more traditional ones including written language or spoken, to corpora of sign language or corpora of video that encode paralinguistic features such as gestures and hyperlinks to video or sound (O'Keeffe and McCarthy 2010; McEnery and Hardie 2012; Ferraresi and Bernardini 2019). The rise of popularity of social media has obviously created a fertile ground for data collection as well. The use of corpora makes CL an "evidence-driven" (Partington, Duguid and Taylor 2013: 5) type of analysis. The most famous and largest, genre-balanced corpus of American English, for example, is the Corpus of Contemporary American English (COCA) (Davies 2008-).[2] This corpus can be defined as a monitor corpus, a corpus that is open and new data is continuously added to it. In this case the compiler continues to update it since 1990, year of its inception. Before COCA, another large corpus of American English was built in the sixties, the Brown Corpus of American English (Francis and Kučera 1964), a ground-breaking project which still today continues to be at the heart of many academic investigations.

Currently corpora, therefore CL, are being used in a variety of fields, from lexicography to language acquisition to discourse analysis (O'Keeffe and McCarthy 2010) but also in connection to other fields such as Literature for example (Mahlberg *et al.* 2019; Culpeper 2009). CL has become a valuable method for bringing together language and language use and spatial patterns in geographical databases (Gregory and Hardie 2011). In Ian Gregory and Andrew Hardie's words, "[m]any branches of the humanities focus on textual evidence; whenever such evidence is considered on the large scale, corpus methods may be of use." (Gregory and Hardie 2011: 298)

[2] Available at https://www.english-corpora.org/coca/.

Let us now turn to DH. Julianne Nyham and Andrew Flinn (2016) claim the field has been in the years been labelled in different ways such as Humanist Informatics, Literary and Linguistic Computing, being the most common Humanities Computing. However, they suggest the term DH started to be widely adopted in 2006. About DH, Nyham and Flinn (2016: 1) posit that DH:

> takes place at the intersection of computing and cultural heritage. It aims to transform how the artefacts (such as manuscripts) and the phenomena (such as attitudes) that the Humanities study can be encountered, transmitted, questioned, interpreted, problematized and imagined. In doing so it tends to differentiate itself from now routine uses of computing in research and teaching, for example, email and word processing.

Additionally, they suggest the increase in the use of the label "Digital Humanities" indicates not only a terminology preference, but also signals the increasing use of digital resources in humanities, emphasizing how for a long time, DH was restricted to research intensive centres that could afford paying for the right equipment, professionals and maintenance.

Some scholars place the start of what can be considered as the DH as back as when Father Roberto Busa, in 1949, right at the end of World War II, initiated the creation of the first digital archive of the works of Saint Thomas Aquinas, the *Index Thomisticus*, a complete lemmatization of the works written by the philosopher (Schreibman *et al*. 2004). This first digital collection of texts has been also viewed as one of the first instances of corpora within CL (see among others Jones 2016). In its most basic definition, we could say that DH is a field of inquiry that brings together the study of humanities through the use of computer-mediated techniques, in other words "using technology to illuminate the human record, and bringing an understanding of the human record to bear on the development and use of information technology" (Schreibman *et al*. 2004: XXIIV). The interdisciplinary core of DH is its most peculiar feature and the one that opens it up to a number of intersections. "[T]he digital humanities reconfigures the humanities for the Internet age, leveraging networked technologies to exchange ideas, create communities of practice, and build knowledge." (Spiro 2012: n.p.) In *Digital Humanities* (2012) Anne Burdick *et al*. present in the first section a description of DH trying to address the question "What are Digital Humanities?" A number of nouns are highlighted in this section which are used

to describe DH, these are: design, computation, digitization, classification, description, metadata, organization, navigation, curation, analysis (processing of text or data), editing, modeling, networks, infrastructure, versioning, prototyping, failures. As a corpus linguist these could have easily been used in an introductory chapter on corpus-based methods for language analysis. Once again, these two worlds intersect and the similarity between the two is more and more apparent.

It must be noted that the intersection between CL and DH has been acknowledged in the past, and references to it can be found in the literature in the past decade. In the book mentioned above (Schreibman *et al*. 2004), for example, Nancy Ide (2004) dedicates a whole chapter to CL and its tools of analysis. Yet, to this day, these fields are generally used as alternatives for one another, and even more so, despite the efforts of CL scholars, the use of these techniques are mostly circumscribed to linguistic studies. In the next section, I will show you how the accessibility and structure of CL appear to be a perfect fit for DH and conversely, the field of DH seems to have room for a rather approachable and relevant methodology such as CL.

Twitter as a Digital Archive: A Case Study on #wontbeerased

Digital archives understood as collections of digital or digitalized language are now easy to find and access and are at the foundation of DH. The growing trend towards digitalization, today more than ever given the semi-worldwide inability to travel and access physical archives due to the COVID-19 health crisis, gave an even more central role to digital archives, which in return became fruitful pools for corpora compilation. As we are surrounded by digital archives and possible sources of data, CL is a great tool to take advantage of them. One of the most commonly used platforms nowadays, Twitter, can be considered as one of these digital archives, a place where language that can constitute a corpus can be found. A collection of language, publicly available, that includes metadata, available digitally and representative of a variety of genres, languages, topics or even personalities. It can serve as a place for data collection not only to analyze language use or language change in time, but to observe political developments, historical events or social behaviors, thus a useful source for many disciplines. Twitter is based on social interaction between users and can be employed to create ambient affiliation through the use of hashtags (Zappavigna 2012).

As Massimiliano Demata (2018) observes, based on the Pew Research Center (2017) data, the intensive use of social media has an impact on how people learn about and understand news and politics. In the USA, for example, 62% of the population accesses news through social media, of these more than half rely on twitter specifically (Demata 2018, 70).

> The use of Twitter as an instrument of political information and propaganda is a relatively new development of the information structure at the basis of a modern democracy. In the last two decades or so, social media have deeply altered the way political information is controlled, distributed and consumed. (Demata 2018: 69)

More generally, because of the "communicative affordance" provided by social media these have been defined as "a new paradigm of communication" (KhosraviNik 2017: 752), used by people more and more.

In this section, I want to discuss a small case study that uses Twitter as a digital archive and looks at the use of language in communication and more specifically discusses the issue of digital activism in the USA context from a linguistic perspective.

The participatory nature of social media has made the Internet a breeding ground for a variety of exclusionary, intolerant, and extremist discourses, practices and beliefs (Kopytowska 2017), especially from politicians who seem to have found in Twitter a very effective means to communicate their thoughts and political ideas. One of the personalities who have found in Twitter a strong ally is U.S. President Donald Trump, as Demata (2018) discusses thoroughly. Trump has used the platform so extensively to the point he has affirmed that he would probably not even be where he is now if it was not for Twitter. The president has become notorious for his controversial tweets used not only for trivial issues but in many cases to announce new political turns and decisions.

The transgender community has been repeatedly targeted by Trump's mediatic communication. One of the latest in his long tradition of suppression of human rights goes back to October 2018. On this occasion, the department of Health and Human services announced it was in the process of revising Title IX of the Federal Civil Rights law to elaborate and establish a legal definition of sex and gender identity that would define gender as a biological and immutable condition[3]. A num-

[3] The source used to retrieve the specific steps in the political agenda of the POTUS is https://transequality.org/the-discrimination-administration.

ber of protests raised across the country to fight this and say loud and clear that transgender people would not be erased, a gathering organized by GLAAD[4] was held in Washington Square Park, NY on the same day as this was announced, followed by a rally in front of the White House in the next days (22/10/2018), for example. At the same time as Twitter was being used by the President of the United States to persecute a minority, this platform became the place for a sort of counterattack and the hashtag wontbeerased became the symbol of this protest. I became interested in the way this *counteratattack* was put forward on Twitter and decided to analyse the linguistic and discursive practices that were being used on the social media to pursue this protest.

The corpus used to carry out the analysis was created by scraping Twitter using an adapted version of a Python library called Get Old Tweets3 (https://pypi.org/project/GetOldTweets3/)[5] the data was then processed into the right input text for the software, i.e. .txt. The data was collected in a time span that stretches across 6 months between October 2018 (when the protest started), until March 2019 (the last full month when the data was collected). Tweets were selected using the hashtag in all its forms as a search term. The search was limited to English language and to the actual post, no replies or retweets were included, generating a corpus of 438,723 tokens. I then resorted to a more CL-based approach, within the framework of Corpus-assisted Discourse Analysis (Partington, Duguid and Taylor 2013), and used the software Wmatrix (Rayson 2009) to analyze the data, a tool best known for its semantic analysis tools.

When data is uploaded to Wmatrix it is automatically tagged by CLAWS (Garside and Smith 1997), a grammatical tagger created at Lancaster University for part-of-speech (each word is provided with information about its grammatical function) and semantically tagged by USAS (Wilson and Rayson 1993) an English semantic tagger also created at Lancaster University. For the purpose of this case study, I focused on the semantic domains generated by the second type of tagging. Simply put, each word is assigned a tag that indicates a semantic domain, i.e. "time" or "emotions". Using log-likelihood[6] as a standard value, a frequency list of the tags is created and compared against the frequency list of the tags of a reference corpus, in this case the reference corpus is BNCwritten sampler (Burnard 1999) available on Wmatrix. From this comparison a final list of the key semantic domains in my corpus is generated. I will now discuss in more detail the three most significant semantic domains in this list. This analysis reveals insights both on the structural use of tweets for activism and about the content of these tweets.

The most significant group in the list of domains (semantic domain tags in the table below) is labelled as "unmatched" (Z99 in the labels used by Wmatrix). There are over 30000 occurrences that match this domain. As the name of the domain suggests, this group contains all the words that the software did not recognize, mostly hashtags and @ signs, which on Twitter are used when you want to reply directly to someone, to tag them in your comment. In other corpora these might be irrelevant, but in the case of a corpus collected from Twitter they are actually key elements.

The table below (Tab. 1) shows the first 20 most frequent hashtags that were found in this domain and their frequency and relative frequency in the corpus.

The centrality, both in terms of number of occurrences and statistical significance[7], of this specific feature of the corpus highlights two main aspects of the way in which activists employed language on Twitter in this case, the first one by means of the use of hashtags and the practice defined by Michele Zappavigna (2018) as hashtagging. Because Twitter only allows a certain amount of words per tweet, the use of hashtags to support the statement being made becomes fundamental. The hashtag acquires the same semiotic function of an image, the meaning is embedded in it, one hashtag carries meaning that you are no longer in need to write and explain, because that simple word already says everything, already brings the meaning with it. The hashtag begins to work in the same way as image, as for example an emoji or meme would, proving once again that online communication is becoming more and more summarized and iconographic. Along the same lines, the use of tagging (through the @ sign) becomes extremely important as it enables direct interaction between the politician – the main actor, and the member of the public who is using the social media to make a claim. In this case, for example, as Table 1 shows, the main interlocutor is

[4] https://www.glaad.org/.

[5] Special thanks to Andressa Rodrigues Gomide for the support in collecting the data presented in this case study.

[6] To read more about log-likelihood and other statistical measures see among others: Evert (2008).

[7] When looking at the table we must bear in mind that we are looking at a very restricted amount of data and that the statistical relevance of the domain is intended in reference to the domain as whole rather than the statistical relevance of the occurrence of each element.

	Word	Semantic domain tag	Frequency	Relative Frequency
1	#wontbeerased	Z99	6067	4.14
2	#transgender	Z99	350	0.60
3	#transisbeautiful	Z99	300	0.24
4	#lgbtq	Z99	233	0.16
5	#transrights	Z99	170	0.12
6	#pride	Z99	153	0.10
7	#wontbeerased.	Z99	149	0.10
8	@realdonaldtrump	Z99	137	0.09
9	#girlslikeus	Z99	127	0.09
10	#resist	Z99	94	0.06
11	#translivesmatter	Z99	92	0.06
12	#loveislove	Z99	91	0.06
13	#nonbinary	Z99	81	0.06
14	#lovewins	Z99	78	0.05
15	#gendertag	Z99	71	0.05
16	#thisisme	Z99	71	0.05
17	#simplerthanwords	Z99	71	0.05
18	#justbeyou	Z99	71	0.05
19	#itgetsbetter	Z99	70	0.05
20	#stopthehate	Z99	67	0.05

TABLE 1: DOMAIN: UNMATCHED.

precisely Donald Trump (Table 1 ex. 8).

The use of the hashtags also allows for the users to launch slogans that become empowering phrases, see for instance examples 11, 14, 17 and 20, not to mention the main hashtag as well, which connote the tweets positively and users as activists.

All in all, from the point of view of the structure of these texts used by the activists we can posit that a short tweet not only enables the users to express a much more complex idea that 280 characters would normally allow, but also opens a direct communication with a specific user. Both of these features would not be allowed by what we can consider as traditional activism, i.e. marches or sit-ins for example. As Sarah Jackson, Moya Bailey and Brooke Foucault Welles (2020: 42) point out, the use of social media and Twitter in particular, has become "one important technology to push the mainstream public sphere on issues of social progress in ways more powerful and visible than possibly ever before."

Moving on to the second and third most relevant key semantic domains we have "people" and "pronouns" which will be discussed together here due to the overlapping content of the two domains. These mainly include collective nouns that refer to human beings and pronouns; 3256 tokens match the "people" domain and over 14 thousand match the "pronouns" domain.

Wmatrix allows the users to access to a concordance list of the words included in that specific semantic domain being analyzed, thus I conducted a concordance analysis on a random sample of 100 concordance lines for both semantic domains. That is to say, I was able to look at the context of use of the terms included in the two domains. The reason why I selected a random sample is related to the fact that my aim here is to have a general overview and not to focus on any specific word included within the domain. These analyses revealed that the tweets are far from being impersonal or generic, but tend to be very personal and above all aim at bringing the discussion back to the people, to underline that it is human beings that are at stake and specify that while this issue is definitely concerned with laws and politics, it should, above all, be discussed in terms of what or whom it truly involves, people's lives that are being questioned. We have two different types of tweets here, which can be differentiated by the level of personal involvement in the tweet itself. In the first case we have tweets that are always written by trans people who use the social media to bring visibility to their identity and their community as whole. An example is the Tweet that can be read below:

> nemowo whats this? :3c @swagsires · 6 nov 2018
>
> #MyTransIs living my best life and standing up for my trans siblings. WontBeErased
>
> : 8/21/18
> Age: 22
> Pronouns: He/Him or They/Them[8].

In this case the user, also through the use of a photographs (which have not been included but can be found in the original tweet) states who he is and declares through the use of the hashtag that he will not be erased by a law that is an explicit political act which does not take into account the lives of thousands.
The second type of tweets are less personal or related to the users' direct personal experience, and most importantly are written both by trans and cisgender people, an example is below.

> Kirsten Gillibrand @SenGillibrand 22 ott 2018
>
> When this administration spews hate, we will speak out louder. When they commit injustice against one of us, we will come together to stand stronger. When they attack our basic human rights, we will fight back harder. Transgender Americans #WontBeErased. We won't stand for it.[9]

Along these lines, one of the most recurring catchphrases that follow the hashtag and that was identified through the concordance analyses, is "trans rights are human rights". The main discourse pattern that is put forward by the Twitter users who employ this hashtag is that this behavior on the part of the POTUS is infringing basic human rights and no matter how big the effort is to erase those words that we use to language gender identity, people still exist, they cannot simply be erased or vanish. This discourse is supported by phrases such as "fighting for rights", "resisting erasure", "deserving to be alive and to exist", "attack our basic human rights" (as shown in the second tweet) which despite recalling battlefield language – i.e. fighting, resisting, present a very positive discourse. The literature in this field has proven more than once that there is a tendency, as anticipated at the beginning of this section, of spreading hate online, using very specific linguistic techniques. Majid KhoshraviNik and Eleonora Esposito (2018), in particular, highlight three specific features to which online haters refer to, and these are anonymity, seen as the ability of the web to hide ones identity, physical separation, which comes as a consequence of anonymity, and it is intended as the practice of distancing oneself from its online identity, the lack of face-to-face interaction and acknowledgement of people's humanity, and lastly the practice of de-individuation which consists in relying on the group, being part of a specific ensemble by reducing self-awareness and self-visibility. The analysis of the small set of data presented here shows that the users, or activists as I have defined them earlier, that used this hashtag use opposite strategies to these, eliminating altogether the practice of anonymity. They do this not only through the words but also through the use of images (see for example the first tweet quoted). The users in this case employ a type of discourse that opposes online hate and that could be defined as online love, where, despite the discourse patterns retrieved might hint at a sort of "battlefield" language, the metaphors are actually used in a positive way and the language is directed towards the production of a beneficial meaning. This positive use of language in these 'battlefield metaphors' recalls the type of positive representation described earlier in this section, where people make use of positive hashtags to accompany their tweets, or in the use of personal comments which can be seen as way to underline the humans behind the social media user account.

The results presented here are very limited and only apply to my dataset, the point, in fact, was not to discuss at length the case study, but to provide an example of the type of analyses that can result from the combination of CL and DH in the framework of American Studies. In fact, this case study proves that the combination of these approaches can answers questions related not only to the use of language on the media or in relation to gender and sexual identities, but can speak to issues such as politics, social theory, activism, the role of social media in society and culture at large as portrayed and put forward in the USA, a country which plays a seminal role worldwide in political and social trends.

Conclusions

In 2014 Jensen affirmed that "CL is on the fringes of contemporary DH, which is itself [...] on the fringes of humanities" (Jensen 2014: 57), six years later is this still the case? DH has grown greatly, as testified by, for instance, the growing offer of degrees and research explicitly labelled as being part of the Digital Humanities not only in the USA, where it first started, but more and more in Europe as well. A testimony of this is also the

[8] Nemowo whats this? :3c, Twitter post, November 2018, 11:26 am, https://twitter.com/swagsires/status/1059768940619067392.
[9] Kirsten Gillibrand, Twitter post, October 2018, 9:56 pm, https://twitter.com/SenGillibrand/status/1054476486793682945.

increasing scholarly publication within the field that aim at the interdisciplinarity of this field that this paper also argues for (see the latest issues of *Digital Humanities Quarterly*). At the same time, CL has also witnessed a growth in number of scholars adopting its methodology, as well areas in which the tools of CL are used, as discussed extensively in previous sections. As Gavin Brookes and tony McEnery (2020: 385) agree, CL and DH "appear to be a good match: both are inextricably tied to digital technology, both use digital or digitised data and both use computational tools for analysis." Their encounter, via other disciplines as well, like in this study, has become not only needed but necessary for scholars of any discipline which involves digital data. For instance thinking about my experience, while some techniques in DH and CL overlap, as the practice of scraping the internet in search of data for example, the limited knowledge that I have when it comes to coding and using script-based techniques would not have allowed me to automatically process the data if it was not for the accessible tools of CL. At the same time, my interest in CL opened up my search for frameworks that are broader and have qualities such as quantitative and open source at their core, which led me to explore DH further. The endeavor of DH "to bring computational methods to humanities research" (Dunst 2016: 381) can be supported and simplified by the tools of CL, and can help overcome those obstacles related to the use of technologies that would otherwise keep many scholars away from DH, as the case study presented in the previous section and my own experience demonstrate.

The similarities between the two approaches, where they both "seek to shed light on one or more aspects of the human experience, and neither is afraid to explore the opportunities offered by digital technology" (Jensen 2014: 131), make DH and CL even more appropriate for one another, and not taking advantage of such richness seems like an enormous loss for academia, in term not only of issues that could be explored, but also in enriching the different areas from knowledge acquired and achieved through the use of an interdisciplinary approach.

Another commonality, as discussed in the introduction to this paper, are the set of values proposed by Spiro (2012) for the DH which include openness, collaboration, collegiality and connectedness, diversity, experimentation, that apply to CL perfectly. While CL tools have until now proved efficient in assisting "linguists to see phenomena and discover patterns which were not previously suspect" (Stubbs 1996: 231), thanks to the combination with DH this could be extended to many more disciplines and fields beyond linguistics.

Crossing the two paths, having the digital world at a fingertip thanks to the accessible tools of CL, would allow the two fields to grow even more, tickling the interest not only of scholars but of learners and the general public as well. In fact, growing effort is being put into the popularization of the use of CL techniques, especially in schools[10] but also in non-educational contexts, for example in the field of translation (Baroni et al. 2006) or product branding.[11]

In this effort being made by CL scholars I see the same one being made by digital humanists.

References

Aarts, Jan and Theo van den Heuvel. 1982. "Corpus-based syntax studies." *Gramma: tijdschrift voor taalkunde in Nijmegen* 7 (2-3): 153–174.

Aarts, Jan and Willem Meijs. 1984. *Corpus linguistics: recent developments in the use of computer corpora in English language research*. Amsterdam: Rodopi.

Arnold, Taylor, Nicolas, Ballier, Paula, Lissón and Lauren Tilton. 2019. "Beyond lexical frequencies: using R for text analysis in the digital humanities". *Language Resources and Evaluation* 53: 707-733.

Baroni, Marco, Kilgarriff, Adam, Pomikálek, Jan and Pavel Rychlý. 2006. "WebBootCaT: instant domain-specific corpora to support human translators." *Proceedings of EAMT. 11th Annual Conference of the European Association for Machine Translation*. Norway, 247–252.

Brinson Curiel, Barbara, Kazanjian, David, Kinney, Katherine, Mailloux, Steven, Mechling, Jay, Rowe, John Carlos, Sánchez, George, Streeby, Shelley and Henry Yu. "Introduction." In *Post-Nationalist American Studies*, edited by John Carlos Rowe, 1-21. Berkeley, Los Angeles: University of California Press.

Bronner, J. Simon. 2012. "American Studies: A Discipline." In *Encyclopedia of America Studies*. Baltimore: Johns Hopkins University Press.

Brooks, Gavin and Tony McEnery. 2020. "Corpus linguistics." In *The Routledge Handbook of English language and Digital Humanities*, edited by Svenja Adolphs and Dawn Knight, 378–404. Oxon, New York: Routledge.

Burdick, Anne, Drucker, Johanna, Lunenfeld, Peter, Presner, Todd and Jeffrey Schnapp. 2012. *Digital Humanities*. Oxford: MIT Press.

[10] See for example the Corpus for Schools project at Lancaster University http://wp.lancs.ac.uk/corpusforschools/project/teaching-materials/.

[11] to read more about this see among others: https://www.sketchengine.eu/user-guide/product-naming-ideas/).

Burnard, Lou. 1999. Users Reference Guide for the BNC Sampler. *The British National Corpus Consortium*. http://www.natcorp.ox.ac.uk/corpus/sampler/.

Culpeper, Jonathan. 2009. "Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet." *International Journal of Corpus Linguistics* 14(1): 29–59.https://doi.org/10.1075/ijcl.14.1.03cul.

Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.

Demata, Massimiliano. 2018. ""I think that maybe I wouldn't be here if it wasn't for Twitter". Donald Trump's Populist Style on Twitter." *Textus* 31 (1), 67-90.

Dunst, Alexander. 2016. "Digital American Studies: An Introduction and Rationale". *Amerikastudien* 61 (3), 381- 395.

Evert, Stefan. 2008. "Corpora and collocations". In *Corpus Linguistics. An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 1212-1248. Berlin: Mouton de Gruyter.

Francis, Nelson, and Henry Kučera. 1964. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for use with Digital Computers*. Department of Linguistics, Brown University, Providence. http://icame.uib.no/brown/bcm.html.

Ferraresi, Adriano and Silvia Bernardini. 2019. "Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings." In *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*, 123–139. Amsterdam, Philadelphia: John Benjamins.

Garside, Roger, and Nicholas Smith. 1997. "A hybrid grammatical tagger: CLAWS4." In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey, Leech, and Anthony McEnery, 102-121. London: Longman.

Gilquin, Gaëtanelle, and Stefan Gries. 2009. "Corpora and experimental methods: A state-of-the-art review." *Corpus Linguistics and Linguistic Theory* 5(1): 1-26 .

Gregory, Ian, and Andrew Hardie. 2011. "Visual GISting: bringing together corpus linguistics and Geographical Information Systems." *Literary and Linguistic Computing* 26 (3): 297–314.

Gries, Stefan. 2009. "What is Corpus Linguistics?" *Language and Linguistics Compass* 3: 1-17.

Gold, Matthew and Lauren Klein. 2016. *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press.

Ide, Nancy. 2004. "Preparation and analysis of linguistic corpora." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens and John Unsworth, 289-305. Oxford: Blackwell.

Izzo, Donatella. 2018. "American Studies in Europe/European American Studies: Local and Global Challenges." *RSA Journal* 29: 184-194

Jackson, Sarah J., Bailey Moya and Brooke Foucault Welles. 2020. *HashtagActivism. Networks of Race and Gender Justice*. Cambridge: The MIT Press.

Jensen, Kim E. 2014. "Linguistics and the Digital Humanities: (Computational) Corpus Linguistics." *Journal of Media and Communication Research* 57: 115-134.

Jones, Steven. 2016. *Roberto Busa, S. J., and the Emergence of Humanities Computing*. London and New York: Routledge.

KhosraviNik, Majid. 2017. "Right Wing Populism in the West: Social Media Discourse and Echo Chambers." *Insight Turkey* 19 (3): 53-68.

KhoshraviNik, Majid, and Eleonora Esposito. 2018. "Online hate, digital discourse and critique: Exploring digitally-mediated discursive e practices of gender-based hostility." *Lodz Papers in Pragmatics* 14(1): 45-68.

Kopytowska, Monika. 2017. "Discourses of hate and radicalism in action". In *Contemporary Discourses of Hate and Radicalism across Space and Genres*, edited by Monika Kopytowska, 1-12. Amsterdam, Philadelphia: John Benjamins.

Levander, Caroline, and Robert Levine (Eds.). 2007. *Hemispheric American Studies*. New Brunswick: Rutgers University Press.

Mahlberg, Michaela, Wiegand, Viola, Stockwell, Peter, and Anthony Hennessey. 2019. "Speech-bundles in the 19th-century English novel." *Language and Literature* 28(4): 326–353.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, Tony, Xiao, Richard and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London, New York: Routledge.

McGillivray, Barbara, Thierry, Poibeau and Pablo Ruiz Fabo. 2020. "Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly* 14(2): 1-11.

Nyhan, Julianne, and Andrew Flinn. 2016. *Computation and the humanities: towards an oral history of digital humanities*. Switzerland: Springer Nature.

O'Keeffe, Anne and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Partington, Alan S., Duguid, Alison and Charlotte Taylor. 2013. *Patterns and Meaning in Discourse. Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Rayson, Paul. 2009. "Wmatrix: a web-based corpus processing environment." Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/.

Santos, Diana. 2019. "Literature studies in Literateca: between digital humanities and corpus linguistics." In *Humanists and the Digital Toolbox: In Honour of Christian-Emil Smith Ore*, edited by Martin Doerr, 89-109. Oslo: Novus Forlag.

Schreibman, Susan, Ray Siemens, and John Unsworth. 2004. *A Companion to Digital Humanities*. Malden Oxford, Victoria: Blackwell Publishing.

Spiro, Lisa. 2012. ""This is why we fight": Defining the Values of the Digital Humanities." In *Debates on the Digital Humanities*, edited by Matthew K. Gold. London, Minneapolis: University of Minnesota Press.

Sprugnoli, Rachele, Gabriella, Pardelli, Federico, Boschetti and Riccardo Del Gratta. 2019. "Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale". *Umanistica Digitale* 5: 59-89.

Stubbs, Michael. 1996. *Text and Corpus Linguistics*. Oxford: Blackwell.

Taylor, Charlotte. 2008. "What is Corpus Linguistics? What the data says." In *ICAME Journal*, 32.

Wilson, Andrew and Paul Rayson. 1993. "Automatic Content Analysis of Spoken Discourse." In *Corpus Based Computational Linguistics*, edited by Clive Souter and Eric Atwell, 215-226. Amsterdam: Rodopi.

Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media: How we Use Language to Create Affiliation on the Web*. London, New York: Continuum.

Zappavigna, Michele. 2018. *Searchable Talk. Hashtags and Social Media Metadiscourse*. London, Oxford, New York, new Delhi, Sidney: Bloomsbury.