Bulletin for Spanish and Portuguese Historical Studies
Journal of the Association for Spanish and Portuguese Historical Studies

Volume 43
Issue 1 *Digital Humanities, under the co-editorship of Andrea R. Davis & Andrew H. Lee*

Article 4

2018

Exploring North-South Identities Using NLP: The Image of Spain in the German Weekly Die Zeit

Elisa Garrido
Carlos III University of Madrid, garrido.elisa@gmail.com

Exploring North-South Identities Using NLP: The Image of Spain in the German Weekly Die Zeit

Cover Page Footnote
This research is possible thanks to UPIER Project, financially supported by the HERA along with the European Commission Joint Research Programme 'Uses of the Past'.

Exploring North-South Identities Using NLP: The Image of Spain in the German Weekly *Die Zeit*

Elisa Garrido

North-South coexistence and the image of southern labor force

In order to construct the community of a nation, governments pursue the affirmation of an identity. The aim of the European Union's cultural policy is to bring out common aspects of Europe's heritage, increasing the sentiment of belonging to a unique yet widely diverse community. This is not just about making culture a full aspect of European action, it also takes cultural matters into account in all its policies and economic issues.

National identity implies a set of policies on language, culture, media, youth and education. After the eighteenth and nineteenth century in Europe, the constructivist school around Anderson and Gellner identified fundamental changes in the political and socio-economic environment when identity politics became the principal instrument the state used to secure its legitimacy.[1] However, the desire to belong to a cultural unity is a mechanism that works equally as a tool for exclusion, as it clearly distinguishes between oneself and others, in-group and out-group.[2] As Walkenhorst pointed out, "the historical-geographical and socio-political characteristics of Europe as a continent have generated a 'multiple identity area' of overlapping territorial and historical spaces at local, regional and national territorial level".[3] Because of these differences, we can consider Europe progressively defined as an *imagined community*. This is a concept developed in 1983 by Anderson in his attempt to explain the origins of nationalism. Anderson pictures every nation as a fictive construction, imagined by societies who perceive themselves as being part of a certain group, motivated by interests to be identified as the same nation, reinforcing themselves and opposing the others.[4] This cultural process is present in every cultural community and is the reason to use most common opposite concepts to speak about *others* in our contemporary life such as north-south, center-periphery, ours-theirs or rational-irrational.

According to John Elliot in *Europe Divided*, the differences between northern and southern European countries began in the complex age of the late

---

[1] Benedict Anderson, *Imagined Communities*, 2nd ed (London: Verso, 2006); Ernest Gellner, *Nations and Nationalism* (Oxford: Blackwell, 1983).

[2] Jűrgen Habermas, *Inclusion of the Other: Studies in Political Theory* (London: Polity Press, 2005), 14, 130, 140, 145.

[3] Heiko Walkenhorst, "Constructing the European Identity - Trap or Gap? European Integration Between Community-Building and Path-Dependency," *Limerick Papers in Politics and Public Administration*, no. 1 (2008): 10.

[4] Anderson, *Imagined Communities*, 170.

sixteenth-century after the conflict between the Protestant North and the Catholic South and the development of those political, social, and religious factors that tended to pull the two types of society apart.[5] The Protestant Reformation started in Germany in the sixteenth Century and the ideas quickly spread to neighboring northern countries while other European regions like Ireland, Italy, Spain and Portugal remained Catholic. Some of them, like Spain and Portugal, were powerful countries in the Colonial Era but this did not ultimately translate into real economic growth. Dainotto, in his study follows this approach and proposes an analysis of the foundation ideas about Europe that continue to define culture, politics, and identity today. He describes the origin of the north-south metaphor of European division, analyzing the theories about a Eurocentrism that stigmatizes its own southern regions. It depicts the cultures of countries such as Greece, Italy, Spain, and Portugal as being irrational and rude in stark comparison to those of the rational, civic-minded nations of northern Europe.[6] Max Weber also assumed that religion played a major role in the development of the economies, arguing that Protestants were more inclined to succeed in business than Catholics.[7] Today, some contemporary studies justify this idea based on the social differences between Protestants and Catholics; certain factors seem to influence the development of the economy and empower the dominance of the Protestant part of Europe. Such factors could include: the secularization and consequent freedom of the economy from religious control, differences in education (self-education by reading the Bible autonomously), the consequences of the Catholic Counter-Reformation and the importance of the Atlantic (slave) trade in creating an autonomous business class that would demand modernizing institutional reform.[8]

In post-war Western Europe, the emergence of the welfare state made economic integration a crucial part of identity politics. The beginnings of the European Union provoked the establishment of a European Economic Community in 1957 and the European Free Trade Association in 1960. The emergence of economic agreements in the Eurozone both increased the need to erase differences and empowered unity, yet the "multiple identity problem" of the European integration process shows the difficulty of forming a collective identity.[9] At the Copenhagen European Summit of 14th and 15th December, 1973, the heads of state or government of the nine member states of the enlarged European Community (Germany, Belgium, France, Italy, Luxemburg, the Netherlands, and the recently added Denmark, Ireland and United Kingdom) affirmed their determination to

[5] John Elliott, *Europe Divided: 1559 – 1598* (London: Wiley-Blackwell, 2000).
[6] Roberto M. Dainotto, *Europe (In Theory)* (Durham: Duke University Press, 2007).
[7] Max Weber, *The Protestant Ethic and the Spirit of Capitalism* (New York: Routledge, 1930).
[8] Cristobal Young, "Religion and Economic Growth in Western Europe: 1500-2000." Paper presented at American Sociological Association Annual Meeting, San Francisco, 2009.
[9] Walkenhorst, "Constructing the European Identity," 13.

introduce the concept of European identity into their common foreign relations.[10] In the preamble to the treaty on the European Union, signed in Maastricht in 1992, it is stated: "European integration must be undertaken with the establishment of the European Communities, creating an ever-closer union among the peoples of Europe"; the intention of making a 'Europe of the peoples' means using culture as a vehicle. Helmut Kohl, from the Christian Democratic Union, who is known as a founder of Modern Europe and served as chancellor from 1982 to 1998, had a strong discourse about Europe being a cultural community. However, after Gerhard Fritz Kurt Schröder, chancellor from 1998 to 2005 and Angela Merkel, current chancellor from 2005, global crisis caused the discourse to become more critical, with certain European partners being accused of a lack of capacity to save and manage money.

A survey carried out in 2010 by the Friedrich Ebert Foundation think-tank suggested that more than 30% of people believed the country was "overrun by foreigners".[11] According to *Der Spiegel,* large sections of the German population are suffering from stress relating to identity. Germans lacking immigration in their own families fear that immigrants could strip their sense of home yet Germans with immigrant backgrounds feel marginalized and foreign. Such a phenomenon today exists alongside another, refugees arriving from a home place which they have just lost.[12]

Guest workers began migrating into Germany as the country's economy gained power during the 1960s. Compared to large colonial nations like Britain or France, the German history of migration is different for two main reasons. Firstly, Germany has never dealt with a post-colonial migration wave of people returning to the motherland. Secondly, Germany experienced the exceptional situation of post-war migration when millions of refugees began to enter the country from areas in the Soviet Union which had been part of the German Reich prior to 1945. After the 1960s, Germany quickly developed the reputation of being a very interesting country for immigrants in particular as it had a potentially strong economy. This attracted labor migrants from southern and south-eastern Europe and the first period of Turkish labor immigration recruitment began in the late 1950s. Following that, migration to Germany increased with the official recruitment of migrant workers from Portugal, Italy, Greece and Spain (PIGS). These worker migrants were known as *Gastarbeiter* or 'guest workers', an expression that precisely defined them and

---

[10] "Declaration on European Identity," *Bulletin of the European Communities,* no. 12 (December 1973): 118–122.

[11] "Merkel Says German Multicultural Society Has Failed," *BBC News*, October 17, 2010, https://www.bbc.com/news/world-europe-11559451

[12] "Germany and Immigration. The Changing Face of the Country," *Der Spiegel*, April 19, 2018, http://www.spiegel.de/international/germany/germany-and-immigration-the-changing-face-of-the-country-a-1203143.html

their status within German society.[13] From the point of view of most of the German population, migrant workers were visitors, useful for filling gaps in the German labor market; they were invited to work under temporary contracts and expected to return home afterwards. Most came from Italy, Greece, Spain, Turkey and Yugoslavia and despite contributing greatly to the German 'economic miracle', were still viewed as temporary guests. Thus, their integration was limited to transitory economic incorporation, as they were not intended to become a permanent part of society.[14]

The guest-worker program of the 1960s and 1970s brought about the transcontinental shift of millions of families, along with their assets, ideals, institutions, languages, music, and food. No one at the Ministry of Labor in 1955 could have imagined the transnational cultures that would soon emerge from this "experiment".[15] The country has been shaped through immigration, since a big part of the country's residents have a migratory background. However, neither German public opinion nor subsequent governments had officially acknowledged that Germany was a country of immigration, historically presented instead as a labor recruiting country.[16] The year 2005 marked a turning point, with the arrival of a controversial immigration law which became an instrument to manage the migration that shaped Germany and its integration.[17] The economic crisis in 2008 meant Germany became a key migrant-receiving country, as some researchers noted[18] and this post-crisis flow of EU migrants from Southern Europe bore a strong resemblance to the post-war guest worker migration.

Using DiaCollo to explore word connections in the German newspaper *Die Zeit*

Today, huge amounts of data circulate on the net that, thanks to cloud technology, can be easily archived and accessed. NLP is a scientific discipline combining the fields of artificial intelligence and linguistics. It is increasingly useful for scholars as data accessibility is becoming one of the pillars of modern scientific culture. DiaCollo is a use case of the CLARIN-D center in the Berlin-

[13] Rita Chin, *The Guest Worker Question in Post-War Germany* (Cambridge: Cambridge University Press, 2007), 47.
[14] Zuzanna Hübschmann, *Migrant Integration Programs: The Case of Germany*, Global Migration Research Paper no. 11 (Geneva: Global Migration Centre, 2015).
[15] Deniz Göktürk et al., *Germany in Transit: Nation and Migration 1955–2005* (Berkeley: University of California Press, 2007), 25.
[16] Chin, *The Guest Worker Question*, 198–199.
[17] R. Ohliger and U. Raiser, *Integration und Migration in Berlin: Zahlen-Daten-Fakten* (Berlin: Beauftragte des Senats von Berlin für Integration und Migration, 2005).
[18] Amanda Klekowski von Koppenfels and Jutta Höhne, "Gastarbeiter Migration Revisited: Consolidating Germany's Position as an Immigration Country," in *South-North Migration of EU Citizens in Times of Crisis*, eds. Jean-Michel Lafleur and Mikolaj Stanek (Berlin: Springer, 2017), 149–174.

Brandenburg Academy of Sciences and Humanities (BBAW) that employ text mining technologies. It is a tool for finding typical word connections — collocations — to a keyword in a certain period of time and delivering a visualized presentation of the results.

What is *collocation*? Text mining is a subfield of data mining, which is, in the same way as NLP, a subfield of artificial intelligence. This interdisciplinary field combines machine learning, statistics and computational linguistics. Text mining technology permits the categorization, extraction and sentiment analysis of large corpora.[19] Collocation is one of many tasks in statistical NLP that consists of a keyword and a *collocator* — a word linked to it, usually in a grammatical or symbolic way. It is an expression consisting of "two or more words that correspond to some conventional way of saying things".[20] Generally, collocations mean two or more words that tend to appear together frequently, such as 'New' and 'York'. Of course, there are many other words that can come after New, such as Zealand, Jersey, etc. so, we look for words that appear frequently with others. Discovering collocations in large text archives can give us surprising results depending on the context; as with many aspects of NLP, context in collocations is very important. For researchers in digital humanities, historical context is everything and that is why tools like DiaCollo could become essential. This software tool, developed in the context of CLARIN, can be used for efficient extraction, comparison, and interactive visualization of collocations from historical text collections in German archives. Why can it be useful? DiaCollo profiles can be used to provide humanities researchers an overview of the discourse topics commonly associated with a particular query term and their variation over time. The association strength depends on the date of their occurrence. The geographical example mentioned before (New York, New Zealand, etc.) is quite easy but, what if we look for a more abstract keyword like "crisis"? In such a case, the collocations and the meaning of the word can be expected to vary widely over time, "reflecting changes in the discourse environment which in the case of a newspaper corpus can themselves be assumed to refer to events in the world at large".[21]

Ludwig Wittgenstein, in his famous definition of meaning in *Philosophical Investigations* said: "the meaning of a word is its usage in the language".[22] DiaCollo gives us the possibility of visualizing the changes in meaning of a keyword that can be traced by means of pair words, i.e., words which appear together in a certain

[19] The term *corpus* refers to a collection of data to be processed and we use the plural *corpora* to speak about more than one corpus.

[20] Christopher Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing* (Cambridge: MIT Press, 1999).

[21] Bryan Jurish, "DiaCollo: On the Trail of Diachronic Collocations," in *CLARIN Annual Conference 2015 Book of Abstracts* (Wroław: Clarin, 2015), 28–31.

[22] Ludwig Wittgenstein, *Philosophical Investigations* (Chicester, UK: Wiley-Blackwell, 2009), § 43.

period of time. In this way, changes in meaning will be directly associated to changes in its particular combinations. These changes in the use of a word can also be interpreted as a sign of, for example, political or cultural change. Over the course of time, the relevance of some connections can fade, and others can become more relevant. DiaCollo makes it possible to observe and compare temporal changes in the frequency of use of groups of these words in context. The software allows exploration of the word connections in a selected time interval (for example, over periods of ten years). In order to be able to track the change of a word and its use, or a group of words over a longer period of time, a large number of corpora are available, where the respective creation or publication time is known and indicated. Corpus archives include the *Deutsches Textarchiv* (2.6K documents, 173M tokens) and a large heterogeneous newspaper corpus (10M documents, 4G tokens). The available texts collections are:

- The German text archive (from 1650 to 2000)
- The digital dictionary of the German language (from 1900 to 2000)
- The weekly newspaper *Die Zeit* (1946-2015)
- A compilation of German / Austrian / Swiss daily and weekly newspapers (1946-2015)

Before looking for results, exploring the corpus we are going to focus our analysis on is necessary. In our research, we are interested in understanding the image of Spain in the context of the European Union and its evolution. We do this by completing qualitative analysis on discourse as a progressive socio-historical formation, characterized by particular ways of using language. The digital analysis text research will give us a general vision of the uses of language to refer to any word linked to Spain from a historical perspective.

We have focused this analysis on *Die Zeit* (1946-2015). This weekly German national newspaper is published in Hamburg in northern Germany. It provides a useful corpus to explore the evolution of *northern thinking* that covers most important European events such as the end of WWII, the unification of Germany, the rise of Europe as a community and the beginning of the economic crisis. *Die Zeit* publishes dossiers, essays, long detailed articles and reports of different authors emphasizing their points of view on a single aspect or topic. It was one of the first German newspapers to be licensed by the British after the war and it played an important role in the resurrection of democracy in West Germany. Its political direction is centrist and liberal or left-liberal and it has been one of the most widely read German weekly newspapers.[23] Marion Dönhoff (1909-2002) was

[23] Peter Humphreys, *Mass Media and Media Policy in Western Europe* (Manchester: Manchester University Press, 1996), 82.

its editor and publisher and one of the most important intellectuals of post-war Germany. She became one of the *Die Zeit* leading columnists and transmitted their ideas writing about politics and foreign policy. In 1946, Dönhoff joined the newspaper as political editor. She was later promoted to deputy editor-in-chief in 1955, then editor-in-chief in 1968 and publisher in 1972. Before her death in 2002, aged 92, Dönhoff was still co-publisher of the influential newspaper. She became the most important journalist in the Federal Republic of Germany and as a long-time editor of *Die Zeit*, made history; becoming a bestselling author and advocate of the reconciliation between East and West. In this sense, she set the moral standards for the coexistence of the peoples in a united Europe.[24] Another main editor was the German chancellor Helmut Schmidt, who served as Chancellor of a left-liberal government (the Social Democratic Party of Germany) for the eight years spanning 1974 to 1982. He was in charge between 1983-2015 and worked alongside the previous editor Dönhoff. H. Schmidt. Schmidt was a pioneer of international economic cooperation and one of the most significant policymakers from the 1960s to the 1980s. As editor of *Die Zeit*, he would become an influential commentator.[25] In the light of this background, we can assume social interests and politics are strongly present in *Die Zeit*.

The exploration of word collocations featured the examples of the language-historical analysis presented within *Die Zeit* in different time periods. There are different profiles we can display in DiaCollo. The results of a simple request are returned as a tabular profile of the *best* collocates for the queried word(s) or phrase(s) in each of the requested date sub-intervals ("epochs" or "slices", e.g. decades) specified by the date and slice parameters. If we leave the date box blank, this means we cover all the publication period stored in DiaCollo: from 1946 to 2015. The time segment is called SLICE (marked 10) and the KBEST (marked 10) is the number of word collocations we want to display. In this analysis, the ten strongest collocations are displayed for each ten year period: a set of ten words for each decade The words common to "Spanien" will be mapped to the respective basic form (Image 1).

[24] Klaus Harpprecht, *Die Gräfin: Marion Dönhoff: eine Biographie* (Hamburg: Rowohlt Verlag, 2008).
[25] Hartmut Soell, *Helmut Schmidt: Pioneer of International Economic and Financial Cooperation* (New York: Springer, 2014).

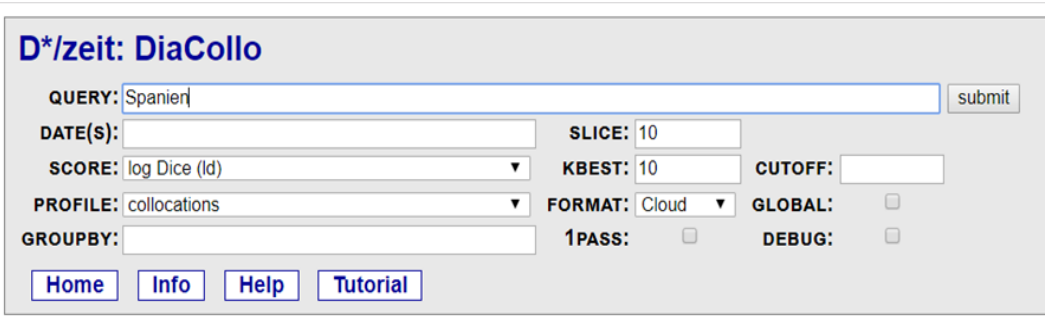

Image 1. DiaCollo interface.

We can choose the word output format from different formats (e.g., HTML, Text, Bubble and Cloud). Next image is showing the Cloud format, which we consider the most visual display, to have a general view of collocations (Image 2). The analysis mode or "profile" in DiaCollo is collocations (a word or phrase that is often used with another word or phrase) so the tool displays the more frequently occurring words in the existing texts.

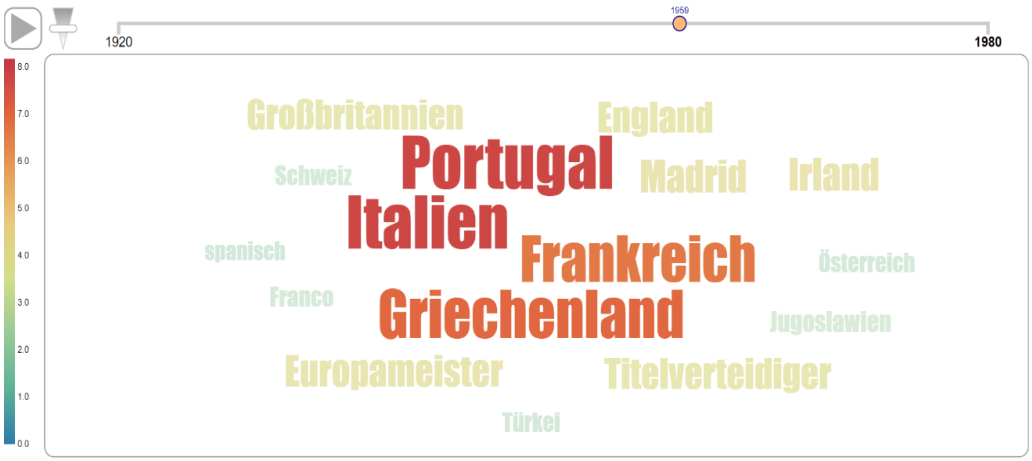

Image 2. Visualization of word collocation results in DiaCollo

The result shows the strength of the displayed collocations with “Spanien”. Dark red indicates a strong collocation (e.g. Portugal and Italy) and the lighter colors like yellow, green and blue, show a weaker link (e.g., England or Turkey). Exploring different time sections, we see some differences in the displayed figures in the specific time selected.

After selecting the HTML format, the application shows the results in a table (Image 3) divided by the period of time selected.

N	f1	f2	f12	score	label	lemma	pos	
15657298	2717	2717	18	6.7621	1940	Spanien	NE	KWIC
15657298	2717	983	11	6.6061	1940	Franco	NE	KWIC
15657298	2717	766	7	6.0412	1940	Portugal	NE	KWIC
15657298	2717	6543	16	5.8232	1940	Beziehung	NN	KWIC
15657298	2717	3130	10	5.8084	1940	Marshall-Plan	NN	KWIC
15657298	2717	18658	33	5.6608	1940	Frankreich	NE	KWIC
15657298	2717	669	5	5.5966	1940	Abbruch	NN	KWIC
15657298	2717	2048	7	5.5891	1940	spanisch	ADJA	KWIC
15657298	2717	5144	11	5.5189	1940	Uno	NN	KWIC
15657298	2717	2471	7	5.4664	1940	Argentinien	NE	KWIC

Image 3. Table in HTML format in DiaCollo.

We use the HTML display to analyze the collocations ranked by importance. After finding chronological word connections, we can trace the results directly to the references and analyze the documents in context, to have a point-accurate analysis. For example, in the 1940s, seeking Spain, we find references to the dictator Francisco Franco, in the context of the end of the Spanish Civil war (1936-1939) and the closed borders of the country (Image 3). In the next decades some popular topics are the international market and the several agreements made after the years of European unification.

The HTML and interactive display formats provide an intuitive color-coded representation of the association score (rsp. score-difference for "diff" profiles) associated with each collocation pair, as well as hyperlinks to underlying corpus hits ("KWIC-links") for each data point displayed. *N* is the total number of co-occurrences in between 'any' two words in the current corpus epoch (date-slice); *f1* is the total independent number of co-occurrences for the query term (w1,*); *f2* is the total independent number of co-occurrences for the collocate term (*,w2); *f12* is the number of co-occurrences for the collocation-pair (w1,w2); and the score is the result of the score functions which is designed to rank those collocates higher which *do* have something "special" to do with the *collocant*. Native collocation profiles retrieve and ranks all content words (w2) occurring together with the search term (w1) within a context window of *d max* content words and without intervening on boundary of the selected DDC break collection. [26] Only collocation pairs with a minimum frequency of *f* are considered. For reasons of efficiency, the frequency threshold *f min* the context-window size *d max*, and the boundary DDC break collection must be specified at compile-time, and cannot be changed by the user.

[26] Jurish, "DiaCollo," 28–31.

The default DiaCollo configuration uses sentence-break boundaries, *d max=5*, and *f min=5*. Supported score functions include absolute raw- and log-frequency (*f, lf*) and normalized raw- and log-frequency per million tokens (*fm, lfm*), pointwise. Candidate collocates are ranked in descending order by associated scores, and the k-best candidates in each epoch are selected and returned. DiaCollo assigns each collocate in a unary profile for a target term score by means of a user-specified score function. The default score is calculated by "logDice ratio", as defined by Rychlý: "reasonable interpretation, scales well on a different corpus size, is stable on subcorpora, and the values are in reasonable range". [27]

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

Image 4. LogDice formula by Rychlý

The log-Dice score function is much more sophisticated than raw frequency. If user ranks by raw frequency (*f12*) alone, it will get a lot of uninteresting candidate collocates ranked at the top (typically function words like determiners), which occur often together with the *collocant* because they occur very often, and not necessarily because they have anything "special" to do with the key word. DiaCollo can filter out function-words, in as much as log-Dice and most of the other score functions, are designed to rank collocations by their relevance.

After analyzing strongest collocations by decade, we note the changes in language by time and we can conclude that the strongest collocations are Portugal, Italy and Greece. This group of countries implies, specifically, the PIGS acronym which refers to the economies of the Southern European countries of Portugal, Italy, Greece, and Spain- those EU members that were less able to refinance their government debt during the crisis. That means that most of articles published under the topic "Spain" in *Die Zeit* from its beginning to now, mention, in some way, Portugal, Italy and Greece, specially from 1960 to 1990 (Table 1).

[27] Pavel Rychlý, "A Lexicographer-Friendly Association Score," in *Proceedings of Recent Advances in Slavonic Natural Language Processing, eds.* P. Sojka, and A. Horák (Karlova Studánka: RASLAN, 2008) 6–9.

1960		1970		1980		1990	
SCORE	LEMMA	SCORE	LEMMA	SCORE	LEMMA	SCORE	LEMMA
7,7888	Portugal	7,7852	Portugal	8,4193	Portugal	8,3312	Portugal
7,386	Italien	7,6921	Griechenland	7,2258	Italien	7,9076	Italien
6,9662	Griechenland	7,0350	Italien	7,2141	Griechenland	7,3032	Griechenland
6,4307	Jugoslawien	6,7498	Jugoslawien	6,4298	Frankreich	6,9751	Frankreich
6,2416	Schweiz	6,3069	Frankreich	6,3588	Beitritt	6,4992	Großbritannien
5,8361	Österreich	6,0714	Türkei	5,5641	Großbritannien	6,4387	Irland
5,732	Franco	5,9079	Franco	5,3948	spanisch	6,2796	Belgien
5,6724	Frankreich	5,4533	Schweiz	5,3528	Franco	5,529	England
5,6362	Ahiers	5,4242	Österreich	5,1422	Österreich	4,3371	Deutchland
5,1877	Türkei	5,3147	spanisch	4,9747	Gemeinschaft	4,1517	Franco

Table 1. Strongest collocations by decade spanning the interval of 1960–1990.

After using corpus linguistic methods to analyze millions of words of news from 1940 to 2010, the results for a unary DiaCollo profile of proper name collocations for the noun Spanien ("Spain"), in 10-year epochs over the archive of *Die Zeit* newspaper, we can visualize the cultural connections among southern European countries.

The Treaty of Rome (1957) determined to "lay the foundations of an ever-closer union among the peoples of Europe" and "eliminate the barriers which divide Europe" was seen as a major stepping stone in the creation of the EU. The Maastricht Treaty, in 1992, had pushed Europeans toward "an even closer union among the peoples of Europe". But, as Dainotto noted, neither Rome nor Maastricht, however, could possibly compare with the news of March 26 in 1995, when seven of the fifteen European Union Members formally dismantled border controls between their countries, setting up the passport free zone and finally, feel part of the "imagined community of other faraway creatures holding, like me, a European passport".[28] Borderless Europe was not only an economic issue but a social phenomenon which affected southern countries' integration with the north. The process brought up the differences between European partners, interacting with such distinct cultures, stressed by the essential differences between the North and the South. After 1995, southern countries like Italy, Portugal, Greece and Spain became officially part of the "southern problem".[29]

It is after 2000, when the economy became a battleground that the topics focus more on the "periphery" of the eurozone. Trying the research with other southern countries (i.e., entering the key-word Portugal, Italy or Greece) produces

[28] Dainotto, *Europe (In Theory)*, 1.
[29] Tanja Borzel, "Why there Is No 'Southern Problem'. On Environmental Leaders and Laggards in the European Union," *Journal of European Public Policy* 7, no. 1 (2000): 141-162.

a similar result of word connections- again showing the neighboring southern countries and re-enforcing the idea of a Southern-identity. Southern European countries were among the hardest hit by the 2008 economic crisis and countries like Germany focused their interest on imposing new austerity policies on Portugal, Italy, Greece and Spain. Consequently, the existing cultural turned into economic differences, employing the PIGS acronym to identify the weaker economies of the eurozone and perpetuating the stereotype of two Europes. Those results doesn´t mean *Die Zeit* presents a negative image of Spain and southern countries, in fact, acronyms like PIGS or PIIGS, that the media has embraced with relish, did not appear in our research. However, there is something the word connection makes visible, an association of these countries- reinforcing the perception of a Europe divided between the core and the periphery.

1990		2000		2010	
SCORE	LEMMA	SCORE	LEMMA	SCORE	LEMMA
8,3312	Portugal	7,9118	Italien	8,2921	Italien
7,9076	Italien	7,6950	Portugal	7,7204	Portugal
7,3032	Griechenland	7,2788	Frankreich	6,9375	Frankreich
6,9751	Frankreich	7,0173	Griechenland	6,776	Griechenland
6,4992	Großbritannien	6,6639	Großbritannien	6,671	Madrid
6,4387	Irland	6,4302	Irland	6,5182	Europameister
6,2796	Belgien	6,1481	England	6,4749	Irland
5,529	England	5,6733	Polen	6,2726	Titelverteidiger
4,3371	Deutchland	5,5898	Belgien	6,166	England
4,1517	Franco	5,4764	José	5,9091	Nadal

Table 2. Strongest collocations by decade spanning the interval of 1990–2010.

Facing this situation, the Spanish press, like the daily newspaper *El País*, complained on multiple occasions about the use of PIGS.[30] During the 1990s, the acronym started to be used in Anglo-Saxon financial press. The *Financial Times* rescued it in 2008 by changing Italy for Ireland to refer to countries with public debt problems, or including the five countries in a new modality: PIIGS. The British newspaper accepted that it was a term full of negativity, although it claimed that there was "a lot of truth" in it. An interesting difference that has affected this idea

[30] Jordi Soler, "Elogio de los PIIGS," *El País*, June 24, 2012, https://elpais.com/elpais/2012/06/22/opinion/1340367240_720933.html; Sandro Pozzi, "El Banco Mundial estigmatiza a los 'PIGS' con una definición errónea," *El País*, June 10, 2010, https://elpais.com/diario/2010/06/10/economia/1276120806_850215.html; Fernando González Laxe, "Entre los BRIC y los PIGS," *El País*, August 30, 2008, https://elpais.com/diario/2008/08/30/galicia/1220091497_850215.html.

and the transmission of the PIGS image over time is the word *Ireland*, which starts to appear in the 1990s for three decades (Table 2). Before the 1990s, Ireland was not included in the acronym but, after the European debt crisis, the term PIGS, (now 'PIIGS') was also increasingly used to refer to the unsuccessful Irish economy (Portugal, Ireland, Italy, Greece and Spain). After 2013, however, with the Irish exit from the eurozone bailout program, PIGS became four again, just as before.[31]

Another possibility of Diacollo is tracking changes in a word's typical collocates over time and provide a clearer picture of diachronic changes in the word's usage. Exploring the concept of "integration" in the German public, discourse can unearth interesting data that can elucidate the situation of Spanish and southern European migrants. DiaCollo brings us the possibility of visualizing in a highchart the evolution of a word through time. Here we can see the evolution of the word "integration" in *Die Zeit*'s life (Image 5).

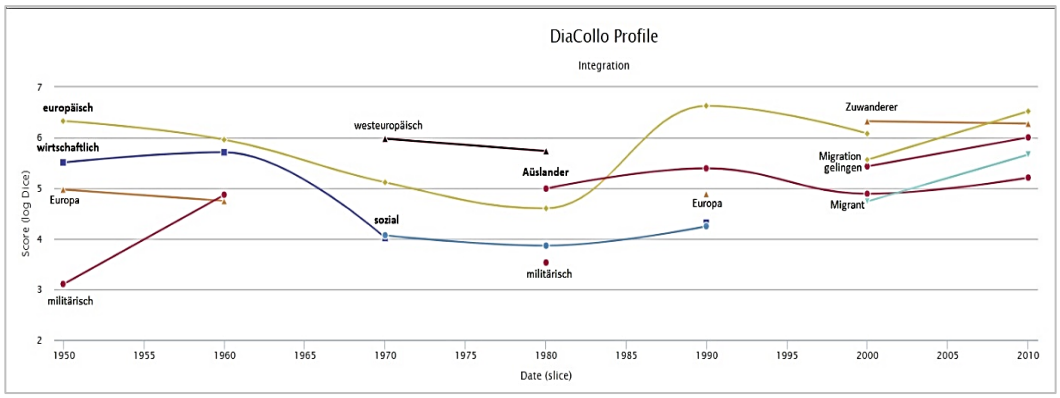

Image 5. Highchart's evolution of the word "integration" through time.

The words that most often collocate with 'integration' are: "*wirtschaftlich*" (economic) and "*europäisch*" (European). From 1950 to 1960, "Europa" (Europe) and "*militärisch*" (military) are some of the popular collocations, which appear again in the 1980s. In the context of postwar migration, German resettlers were fleeing discrimination and persecution from the former communist "Eastern bloc". The importance of military forces could also be explained by the US rearmament program (*Wiederbewaffnung*), the American program to help build up the West German military.

[31] Other acronym with the same pejorative intention was GIPSY (referring to Greece, Ireland, Portugal, Spain and Italy).

Between 1955 and 1968 West Germany signed labor recruitment treaties with Italy, Spain, Greece, Turkey, Morocco, Portugal and Tunisia. The guiding principle of the migration policy for guest workers in the 1950s and 1960s was a "rotation principle" and migrant guest workers were expected to return to their countries of origin once they were no longer needed. [32] Guest workers were encouraged to return to their countries of origin, but most decided to stay and applied for visas for themselves and their families. The *Act on Foreigners* of 1965 regulated entry into Germany as well as the residence status of foreigners. In 1969, West Germany enacted the Law on European Economic Community (EEC) Residence to implement European Community (EC) law regarding the freedom of movement for workers from EEC Member States. It was after the 1970s when West Germany first acknowledged the continuing presence of a large number of migrants in the country and started a formal policy of integration. In graphics, "sozial" (social) appears in 1970 as a collocation and remains until the 1990s. The concept *social integration* was present in the politic discourse about immigration, although CDU repudiated any effort to transform guest workers into Germans and encouraged foreigners to preserve their national and cultural identifications so that they might eventually return to their homelands.[33]

In the 1980s the collocation "*Aüslander*" (foreigners) appears and remains until 2000s. It was in 1983, when West Germany enacted the so-called *Return Assistance Act*, an attempt to encourage guest workers to return to their country of origin. Another *Act on Foreigners* in 1990 regulated entry into Germany and after this, the residence status of foreigners mainly continued from the previous policy but with new rules on legal rights for family reunification and naturalization for the second-generation. Starting in the 2000s, concepts like "*Zuwanderer*" (immigrant), "migration" and "migrant" become popular terms in the German public discourse associated to their integration. After the 2000s the government set up the Independent Commission Migration, which published its report in July 2001 affirming Germany as an "immigration country".[34] It is notable that another word began to be used in articles talking about "integration", that word being "*gelingen*" (successful). The results show an increasing use of "successful" during the decade after the application of the first German Immigration Law in 2005, when, for the first time, the focus was placed on integration measures for long-term, permanent migrant-residents.

From 2016 (not present in graphics here), the so-called Integration Act - *Integrationsgesetz*- promoted the rapid integration of foreigners into the labor

---

[32] Jenny Gesley. *Germany: The Development of Migration and Citizenship Law in Post-War Germany*. (Washington, DC: Global Legal Research Center, 2017), https://www.loc.gov/law/help/migration-citizenship/germany.php

[33] Chin, *The Guest Worker Question*, 98.

[34] Gesley, *Germany*, 9.

market. Under the motto "together we are strong", the support for integration developed especially for the arrival of refugees included German language courses and a guide for asking for permanent residence in Germany; migrants and their families were regarded as a dispensable issue.

Conclusions and future research

We have seen how NLP and collocations can be useful, but clearly, information with context and complementary research needs to be gained. We can conclude that humanists can use text analysis, visualization and mining tools for research to get relevant information initially and that this information can then be relevant in helping them make choices. An important part of the information we gained with NLP was to analyze these resources to understand other opinions, evaluations and emotions, and that part needs further investigation. One possibility offered by DiaCollo is to explore the Key Word in Context (KWIC), that means, after the results of word collocation, the researcher can access the articles statistically ranked, read them and extract all the information written. In a critical sense, following a Foucauldian idea, discourse analysis by corpus can contribute not only to quantitative studies but also to discussion of the interpretation of words in a particular context, as a complex relationship between language, ideology and society. The change from the structuralist concept of language to a more complex post-structural abstract system emphasizes the process of creating ideas and power structures, making discourse a useful tool for understanding historical and social constructions of our era.[35] One of our challenges in this kind of research is manipulation. A solid work of historical contextualization to recognize oriented information and biased material is necessary. In conclusion, NLP and text mining offers unlimited possibilities but also has important limitations. In every case, the result must be understood as an incomplete and partial work that is always susceptible to change. Such affirmations are applicable also to the results revealed in this paper where the information obtained from a German newspaper, whilst not necessarily a universal example of the general view of the northern societies of Europe, can be used to form an initial idea of it. After that, critical research can be developed after the initial information is combined with other resources.

Discourse is a practice that obeys rules (practices, technologies, locations), so despite the fact that corpus analysis tools like DiaCollo give us the possibility of working with empirical data, we need to accept the emergence of giving a new focus on qualitative analysis. This should occur with a view of discourse as being part of the socio-historic process of cultural definition, characterized by ways of using language as a result of identity, ideas and meanings. NLP is a broad concept that can be applied to many aspects of linguistic use. Is it possible, for example, to

[35] Michael Foucault, *The Archaeology of Knowledge* (London: Tavistock, 1972) and *Power/Knowledge* (Brighton: Harvester, 1980).

identify social expression of feelings present in linguistic production, exploring language in corpus by text analysis tools? As texts offer not only simple data but also affective information, such as expressions of emotion,[36] for researchers in humanities, natural language processing (NLP) could be an effective tool of research. An interesting point is the development of sentiment analysis, a way to deeply analyze definitions and interpretations of the topic we are working on. The development of these technologies should give us information about opinions or sentiments but may also create some problems.[37] One of them is determining which documents are relevant, difficult in historical research since our online data are still limited. The challenge of sentiment analysis is in dealing with the computational treatment of opinion, sentiment, and subjectivity in text, which still needs human intervention to be deeply explored and understood. Today, more and more people are making their opinions available to the world via the Internet yet for historical research, sentiment analysis becomes complicated. These tools are undoubtedly useful, but results should not be considered as conclusive, they should instead be reviewed in the background of a dialogue around human assisted interpretation.

Another critical issue is that researchers in humanities are not usually able to use this kind of tool correctly since they cannot fully understand it. The terminology used can be very confusing if the researcher is not familiar with mathematical and statistical data. The articles are usually very technical even though the method can be easy to explain and the combination of infinite possibilities can also cause problems. In DiaCollo, the scoring function calculates the top words that appear near the target term (in this case we have been working with the words *Spain* and *Integration*) and we can use multiple functions to conclude results for statistical significance from the raw counts. To explain how DiaCollo works, the application's designer has worked hard to provide enough papers[38] with prose descriptions of formulation and computational processes, but non-technical readers still could be intimidated by such numbers and formulae.[39] One of the most important issues is the default scoring function which for DiaCollo is "scaled log-Dice ratio". As explained by the application designer, since raw frequency alone is often not a good indicator of association strength, according to Evert[40] each candidate collocate is assigned a scalar association score by means of

[36] B. Liu, *Sentiment Analysis: Mining Opinion, Sentiments and Emotions* (New York: Cambridge University Press, 2015).
[37] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, nos. 1–2 (2008): 1-135.
[38] I want to thank Bryan Jurish for his assistance and invaluable guidance.
[39] Bryan Jurish, "Diachronic Collocations, Genre, and DiaCollo," in *Diachronic Corpora, Genre, and Language Change*, ed. R. J. Whitt (Amsterdam: John Benjamins, 2018) 42–64.
[40] Stefan Evert, *The Statistics of Word Cooccurrences: Word Pairs and Collocations* (PhD diss., Universität Stuttgart, 2005).

the quaternary operation specified by the 'score' (ϕ) parameter.[41] After reading the prose, it is easy to understand why log Dice is the default score, as it seems to be more efficient in getting specific results. Yet, the difficulties historians have in interpreting and assessing such statistical methods are crucial because this can totally change the results. If "log-Dice ratio" is changed to a different score, DiaCollo returns disparate results and it is a very difficult task for a non-technical professional to fully understand such mathematical transformation, its consequences and meanings.

In conclusion, digital humanities are the result of a complex convergence which needs to distribute concepts, categories and objects, as well as associated practices, all in a new context. Technical information needs to address a broader public and humanities should be part of any scientific career. Rapid changes have been marked by the invasion of computer technology into every aspect of life, creating a need for a herculean effort by communities to collaborate in the new discourses and procedures. As stated before, machine learning needs human intervention yet certainly *humans need other humans* to fully succeed.

[41] Jurish, "Diachronic Collocations," 42–64.