https://dx.doi.org/10.4324/9780429425899

QUERY COMPLEXITY AND QUERY REFINEMENT: USING WEB SEARCH FROM A CORPUS PERSPECTIVE WITH DIGITAL NATIVES

Maristella Gatto – Università degli Studi di Bari "Aldo Moro" (Italy)

Abstract

At the dawn of the new Millennium, Prensky (2001a; 2001b) popularized a view of the younger generations as "digital natives". While this myth has not gone unchallenged (Prensky 2009; Thomas 2011), students have since then often been acritically assumed to be almost naturally fluent ICT users. However, recent studies have revealed that this is not always the case. As far as web search in particular is concerned, a tendency has been shown towards a simplistic "get in, get the answer, get out" approach (Thompson 2013: 20-21) which prevents them from taking full advantage of the web's potential for autonomous learning. In this context, this chapter advocates the importance of familiarizing the students with advanced web search as an opportunity for a rewarding accessible DDL experience, thus also contributing to their development of critical reasoning skills.

Introduction

Since the emergence of the notion of the "digital natives", derived from publications by Tapscott (1998) and Prensky (2001a; 2001b) and further supported by a range of other popular appropriations of the term, younger students have been often assumed to possess knowledge and skills that should allow them to handle ICT tools in a natural, fluent way. The very fact that younger people's lives appear to be saturated with digital media has led to the claim that "digital natives" (roughly identified as those born after 1980) might have developed different learning styles and behaviours, in terms of abilities, preferences, attitudes, and even "productiveness" (i.e. focused attention, deep processing, and persistence), precisely as a consequence of their virtually total immersion in digital technology since early childhood and during adolescence (Thompson 2013: 12). However, this assumption has not gone unchallenged. Indeed, ICT ownership and experiences, as well as confidence with ICT devices, do not necessarily imply competent use, and the overall conclusion of many recent studies is that digital natives are not necessarily ICT literates. On the contrary, it is advocated that information literacy should be explicitly enhanced with hands-on and minds-on courses (Šorgo et al. 2016).

As far as web search in particular is concerned, all models for Information Retrieval emphasize the dynamic nature of the search process, suggesting that users learn from their searches, and that their information needs can be adjusted on the basis of retrieval results (Baeza-Yates & Ribeiro-Neto

2011: 23). As suggested in Deschryver and Spiro (2009: 4), the ideal method of learning from the web is an iterative search process, where learners "create their own search phrases based on information they encounter on the Web". However, a tendency has been shown in digital natives towards "fast, expedient web search", which suggests that they mostly adopt a simplistic "get in, get the answer, get out" approach (Thompson 2013: 20-21). Such students may not be taking full advantage of the immense potential of web search, unless they learn to go beyond simple searches and exploit the full the affordances of search engines.

It is against this background that this chapter aims to contribute to the debate on the potential of web searching for Data-Driven Learning. In particular, building on existing literature on both the web as corpus and on so-called Google-Assisted Language Learning (Chinnery 2008; Gatto 2009 and 2014; Eu 2017), as well as on recent studies more specifically focused on uses of web search in DDL (Boulton 2015), I claim that using the web as a corpus through ordinary search engines necessarily engages the learner in a process of progressive query refinement towards greater complexity, which can pave the way to a subsequent appreciation of dedicated corpus linguistics resources and tools. I discuss general issues concerning the learning styles of digital natives with a special focus on web search as a pervasive paradigm in their everyday life, before reconsidering key issues concerning the web as corpus debate and the use of web search from a corpus linguistics perspective. I then review existing literature on DDL using the web for pedagogical purposes, before reporting on a classroom experience based on the use of web search from a DDL perspective with Italian secondary school EFL students. More specifically, the classroom experience is focused on language learning activities in which the students were asked to use the web to gather evidence of language usage, explore phraseology, and test translation candidates. While using the web for activities of this kind is increasingly seen as an appropriate DDL experience (Boulton 2015), it is self-evident that the anarchy of the web, and the resulting problems in locating relevant and reliable results (Gatto 2009) may impair the value of the experience as a whole. The study aims to show the pitfalls of a naïve use of web search for linguistic purposes by so-called "digital natives", and claims that it is only when introduced to the principles of progressive query refinement, in which the web is seen from the perspective of corpus linguistics, that the web can really become the right place for a rewarding DDL experience. The basic assumption of the chapter as a whole is that far from being naturally 'fluent' users of technology, digital natives can still benefit much by becoming familiar with basic notions of information retrieval for linguistic purposes, and that using the web as a corpus in DDL activities can contribute to developing their general critical reasoning skills.

Deconstructing the myth: from digital native to digital wise?

Since Prensky's vision of younger generations as "digital natives" (2001a; 2001b), such students have been thought to have developed a distinctive set of characteristics including "preference for speed, nonlinear processing, multitasking, and social learning" (Thompson 2013: 12), in a way which apparently marks a gap with previous generations. Indeed, as noted in Thomas (2011: 6) the discourse of digital natives is roughly based on three main assumptions in which younger generations are said to:

1. constitute a largely homogenous generation and speak a different language vis-à-vis digital technologies, as opposed to their parents, the "Digital Immigrants";
2. learn differently from preceding generations of students;
3. demand new ways of teaching and learning involving technology.

This characterization apparently assumes, especially in a teaching context, that younger people spontaneously know everything they need to know about technology, rather than having to make the effort to learn about it. In fact, while some correlation has been shown between confident use of digital technology and characteristics generally ascribed to digital native learners, ongoing research on such a correlation seems to suggest a less deterministic relationship between technology and learning. Recent research has shown that so-called digital natives are not necessarily information literate, and that educators should first introduce basic information literacy skills with "hands-on and minds-on courses" (Šorgo 2016). It has also been argued that it is parents and teachers that often construct their children/pupils as ICT experts (Buckingham 2011: ix) without actually testing such supposed experience.

Indeed, when addressing the digital literacy skills of learners and the resulting necessary evolution of forms of pedagogy, teachers should rather be wary of adopting sharp dichotomies based on the binary logic of "natives" and "immigrants". Prensky himself, in his contribution to the book *Deconstructing Digital Natives* (2011), considers how important it is to reflect on the way the metaphor of the 'native' in the digital world has actually entered the popular imagination in ways he could "neither have imagined nor foreseen" (Prensky 2011: 15). The metaphor of the native had been devised as an image for conceptualizing the differences that could be observed in the attitudes of younger and older people when faced with digital technology. While proving a useful evocative metaphor, the image was eventually misunderstood as an "absurd claim […] that if someone was born after a certain date, and was therefore included as a Digital Native, that person automatically knew everything there was to know about digital technology" (Prensky 2011: 27). As Prensky himself maintains, "having grown up with digital technology as toys, Digital Natives are much more at ease with its use than the generation that did not. But this surely doesn't mean they know

everything, or even want to" (Prensky 2011: 27). It is precisely this gap between supposed or assumed fluency and actual knowledge and competence that has made Prensky revise his concept of digital nativeness in terms of "digital wisdom" (Prensky 2011: 30). Digital wisdom, according to Prensky, is a twofold concept which encompasses the "wisdom arising from the use of digital technology to access cognitive power beyond our innate capacity" and the "wisdom in the prudent use of technology to enhance our capabilities". Technology alone, he argues, "will not replace intuition, good judgment, problem-solving abilities, and a clear moral compass" (Prensky 2011: 18).

It is against this complex background that views about the supposed technological fluency (Prensky 2001a:1) by digital natives are now challenged. While the use of digital technology for basic communication seems to be most common among digital natives, very few engage in more complex activities, and there seems to be evidence of a restricted range of technologies centred mostly on mobile phone features and basic web use (e.g., sending an email or looking up information). Many so-called digital natives are no more intensive users of digital media than many adult digital immigrants and young people's use of digital technology can be defined as "mundane rather than spectacular", i.e. characterized not by dramatic manifestations of innovation and creativity, but by relatively routine forms of communication and information retrieval (Buckingham 2011: x).

The web as corpus: **web search from a corpus perspective**

While it may be necessary to tone down enthusiastic claims about digital natives being *ipso facto* ICT experts, it cannot be denied that younger generations are very likely to be familiar with basic ICT activities such as web search, which can be profitably used in language learning.

Web search in is ranked very high among the ICT skills which so-called digital natives are credited with (Thompson, 2013). Nonetheless, research on actual information-seeking performance has shown that information-seeking behaviour by digital natives reflects "a culture of "satisficing" decision-making that is in turn indicative of a surface approach to learning" (Kennedy & Judd 2011: 240). Students often adopt a style of "fast, expedient web search, rather than a more iterative style of searching for information on the web" (Thompson 2013: 21), a finding which is consistent with claims about digital natives being used to "twitch speed" (Prensky 2001b: 442) and accustomed to gathering information quickly, thus lending additional support to the widely held view that learners, whatever their age, might not yet be taking full advantage of the affordances of the web for learning (Thompson 2013).

Indeed, online information seeking – performed through increasingly powerful and efficient search engines - has a great potential to support deeper learning, provided that students go beyond the simplistic "get in, get the answer, get out" approach mentioned above. The ideal method for learning from the Web has been described by Deschryver and Spiro (2009: 4) as an iterative search process, where learners "create their own search phrases based on information they encounter on the Web, either by employing specific ideas from the current page as new search phrases, or by conceptualizing novel search phrases based on recent activity, past experience, and/or the related momentum web learning affords". As argued in Thompson (2013), the distance between this ideal process and the "in and out" approach which seems to be common among younger learners, suggests that the digital natives' supposedly innate confidence with technology needs to be complemented with explicit instruction in such skills as crafting and refining their query as well as evaluating the results, in order to make effective use of the technology that surrounds their lives.

It is in this context that greater attention is being paid to the potential of web search as a gateway to DDL and corpus linguistics. Despite obvious differences between proper corpus linguistics and using the web as a ready-made corpus to be accessed via ordinary web search engines (Gatto 2009; Boulton 2015), the web is by now an accepted language resource with great potential as a source of evidence of attested authentic language use, and the web can well be said to provide linguists (and learners) with "a fabulous linguists' playground" (Kilgarriff and Grefenstette 2003: 345).

Nonetheless, the way to treating the web as a linguistic corpus is by no means straightforward. A comprehensive discussion of the differences between the web as a spontaneous, self-generating collection of texts, and a corpus, defined as a finite-sized *body* of authentic text sampled so as to be representative of a language variety (McEnery & Wilson 2001: 32), definitely falls beyond the scope of the present chapter. My own research (Gatto, 2009; 2010; 2014) has discussed a range of issues with the web as corpus approach, including authenticity, size, representativeness, balance, and verifiability. Among them, authenticity and size have notable consequences for the potential of web as corpus approaches and methods in the context of DDL. As the most prominent feature of the web to have attracted linguists' attention, the web's undisputable nature as a reservoir of authentic texts produced in authentic human interactions is one of its most obvious strengths when considered as a corpus. However, this is also one of its major flaws, and the reason for being particularly cautious when using web search for linguistic purposes. Owing to the web's nature as an unplanned unsupervised unedited collection of texts, authenticity is often problematic. Everyday experience suggests that 'authentic' in the web often means 'inaccurate' (misspelt words, grammar mistakes,

improper usage by non-native speakers). As a consequence, it is of crucial importance that anyone purporting to look at the web from a corpus linguistics perspective, especially for pedagogical purposes, needs to become familiar with some of its basic features so that they can profit from its potential without running the risk of being tangled in the web itself.

Aside from authenticity, when frequency data obtained from the web are to be used as evidence of attested usage, it is of fundamental importance for learners to be aware of the impact of size. In most cases, even a very large number of hits for a given query (especially for languages like English) does not necessarily imply uncontroversial validation of a hypothesis. A case in point in this respect is the example of the unusual collocation "suggestive landscapes", as discussed in Gatto (2009), where frequency data from the web seems to provide abundant evidence of attested usage. The problem, then, concerns the correct interpretation of these results, not only from a quantitative point of view (a task virtually impossible given the immeasurable nature of the web), but also from a qualitative point of view. It requires a closer inspection into the results to probe their reliability, which - in this case - is called into question by the provenance of the results mostly from Italy or from sites referring to Italy (and possibly resulting, therefore, from a literal translation of the Italian "paesaggi suggestivi"). As a matter of fact, the number of hits for this query dramatically drops when resorting to the options "Pages in English" and "Pages from the UK", aimed to retrieve only English pages coming from the United Kingdom.

Two more crucial issues among those discussed in the web as corpus literature deserve special attention in this context, namely relevance and reliability (Lüdeling et al. 2007; Gatto 2009), which relate to two aspects referred to in information retrieval as precision and recall. Any linguistic search carried out by means of specific software tools on a traditional corpus of finite size (such as the BNC) would report 'only' results exactly matching the query (precision), and 'all' the results matching the query (recall), whereas this is patently not so with the web. This can be exemplified by a search aimed at evaluating "onset site" as a translation equivalent for "sede di insorgenza" to be used in a medical text about cancer, as discussed in Gatto (2011). By searching the exact phrase "onset site", the results returned by the search engine would provide a generic quantitative indication about its frequency of occurrence in the web, which roughly gives an indication of appropriateness. In order to increase the relevance of the results to the domain of interest, one might include the word "cancer" in the query so as to retrieve only texts having the word "cancer" somewhere on the page, and therefore more likely to be addressing this specific topic. The lower number of matches for the query *"onset site" cancer* when compared to *"site of onset" cancer*

suggests that “onset site” may not be the preferred wording in this context. Moreover, it should be noted that most of the occurrences reported for the query *“onset site” cancer* feature the two nouns *onset* and *site* separated by punctuation marks, generally a comma, which suggests that these results cannot be considered as evidence of attested usage for the noun phrase “onset site” in this specific context. At this stage, further restriction of the query to a known domain such as a portal for the online distribution of scientific journals (e.g. Elsevier or Pubmed), or simply to Google Books, can be attempted in order to boost the reliability of the results. This finally confirms the inappropriateness of “onset site” in this context, whereas the alternative search for *“site of onset” cancer* within the same restricted domains still produces a significant number of matches, all to be considered relevant and reliable. This example reveals that even the apparently trivial task of searching the web for evidence of usage poses specific problems for the researcher, which require that cautionary procedures are adopted both in the interpretation of the results and in submitting the query to the search engine.

More precisely, it can be argued that the query is the place where the practice of web search meets a linguistically-oriented approach to the web as a corpus. Indeed, using the web as a linguistic resource can be viewed as a specific case of information retrieval. The classic model, as reported in Baeza-Yates & Ribeiro-Neto (2011: 23ff) is formulated as a cycle consisting of four main activities:

- problem identification,
- articulation of information need(s),
- query formulation, and
- results evaluation.

Given the dynamic nature of the search process it is also stressed that users learn from their search, and that their information needs can be adjusted as they see retrieval results (Baeza-Yates & Robeiro-Neto 2011: 23). Using the web as a corpus through ordinary search engines therefore engages the user in a process of progressive query refinement towards greater complexity which can only be supported by greater familiarity with the advanced search options provided by most search engines. As the examples provided have shown, each operator in advanced web search can be interpreted from a corpus linguistics perspective. Thus, while a search for a single word can be compared to the creation of a sort of sub-corpus (made up of all the pages and only the pages containing that word), the search for phrases, possibly combined with the use of wildcards, can represent the search for collocates or patterns. Finally, language, URL and domain restrictions, or search in specific subsections of the web, can indirectly be read in terms of constraints at the level of geographical variation, register and genre (Gatto 2009: 71-77). It is therefore only through a

process of progressive query refinement towards greater complexity that the common practice of web search and the linguist's approach to the web as a corpus most fruitfully interact.

Using web search in Data-Driven Learning

Despite some *caveats*, the practice of using the web as a corpus through ordinary search engines has received great attention in the context of Computer Aided Language Learning (Shei 2008; Wu et al 2009; Sha, 2010; Geluso 2011), sometimes even labelled as Google Assisted Language Learning (Chinnery 2008). In all these studies, while the limitations of the web as a source of language data are invariably acknowledged, its usefulness as a ready-made source of authentic language available at the click of a mouse button is undisputed, especially when one is to test personal intuitions about language use on the basis of existing data. Thanks to its nature as a low tech, low-cost and ubiquitously available option, the web often represents an immediate source of corpus-like linguistic information which makes it a perfect choice in teaching contexts at any level whenever there are limitations for the introduction of more traditional forms of DDL. Indeed, while attempts have been made over the past few years at making tools and resources more and more accessible – both in terms of availability for free and in terms of user-friendliness – barriers still exist that prevent a breakthrough of DDL in the classroom (Boulton 2015: 267; Schaeffer-Lacroix, this volume). To say the least, corpus-based data-driven learning activities still require some comfort and confidence with the technology by teachers and students alike, which cannot always be taken for granted. By contrast, using the web as a corpus surrogate (Baroni & Bernardini 2006) with search engines as surrogate concordancers, seems to provide a valuable opportunity to explore the undeniable *de facto* intersection between web search as a common everyday practice, and linguistically-oriented web search informed by the corpus linguistics approach (Gatto 2009; Boulton 2015). It cannot be denied in fact that the very act of searching the web via an ordinary search engine and reading vertically through the results, as it typically happens even in the simplest web search, is strikingly similar to what happens when searching a corpus through a concordancer. Furthermore, the results page resembles a concordance list where the search item is highlighted within a small amount of co-text, even though this is by no means the same format as the Key Word in Context (KWIC) typical of linguistically oriented tools. It is as if reading "vertically", "fragmented", and looking for "repeated events", which Tognini Bonelli singles out as features that set apart the act of reading through a corpus from the common act of reading a text (2001: 3), is now everyday experience for people beyond the corpus linguistics community, including younger learners. Furthermore, as convincingly argued in Boulton (2015: 268), it should be borne in mind

that DDL does not necessarily rely essentially on a corpus as it is generally understood in corpus linguistics: what is important is that the language should be pedagogically relevant (cf. Braun 2005), and that the learner should engage directly with the data rather than relying on the teacher as intermediary (Boulton 2015: 267).

As a matter of fact, many students acknowledge using the web for linguistic purposes, in many different ways. However, when it comes to using the web as source of attested usage, they eventually display a limited awareness of the real workings of a search engine, which is fundamental knowledge if one is to avoid the common pitfalls and fully exploit the web's potential. Certainly, search engines are designed neither for linguistic purposes nor for data-driven learning, as they aim, in fact, to answer a variety of other needs. Furthermore, it should be observed that the very strategies that enhance the effectiveness of search engines for general purposes, such as normalization of spelling and lemmatization, or the inclusion of synonyms in the results, might fight against precision in a linguistically-oriented web search. And still, these limitations are only the tip of the iceberg. Today it should also be stressed that search engines are engaged in an unprecedented effort to shift the original query-based retrieval algorithm, in which the search engine's task was to find items exactly matching the query into a context-driven information supply algorithm, whose aim is to provide information that is supposed to be of interest to the user on the basis of elements not necessarily contained in the query itself (Broder 2006; Gatto 2014). It would be therefore an oversimplification to think that the problem posed by search engines for data driven learning is that they are not designed for linguists and that they are 'only' geared towards the retrieval of general information from the web. The real problem is that the very needs which search engines address are constantly evolving, and next-generation search engines do not simply 'passively' retrieve the information required but rather 'actively' supply (unsolicited and often commercial) information, both in the form of banner ads or sponsored links, but also through algorithms aimed at behavioural and contextual targeting (Levene 2010: 152 ff.). These new algorithms clearly tend towards strong customization of the results, creating a "filter bubble" (Parser 2011) around each individual user, without the user even being aware of what is happening. This is the reason why most general-purpose web search is now extremely effective and generally successful, even without particular effort on the searcher's part. On the contrary, linguistically-oriented web search, which is quintessentially informational, needs to be grounded on greater awareness of both the nature of the web as a surrogate corpus and of the real workings of search engines as surrogate corpus tools. Hence a deeper knowledge of advanced web search options has long been advocated as the only way to profitably use the web as a corpus especially for DDL.

Among the various operators, such as inclusion/exclusion of context words to limit irrelevant hits for polysemous words, restriction by language, file type, domain (e.g. .ac.uk or .edu for UK and American academic websites) and provenance of the results, the phrase search has been often singled out as a good starting point for entry into the web as corpus, in particular by Maniez (2007), Shei (2008), Acar et al. (2011) and Geluso (2013). The next section focuses on the phrase search as a useful first strategy for younger learners negotiating the data-driven validation of collocates and for testing translation candidates through increasingly complex and refined queries.

Query complexity and query refinement. A classroom experience with digital natives

The approach to DDL used for this experience was mostly derived from research on the web as corpus, with a special focus on the use of frequency data from the web as evidence of attested usage for the validation of collocations and translation candidates. Emphasis was also placed on issues concerning the evaluation of the relevance and reliability of the results and on strategies to enhance these two crucial aspects in information retrieval. More specific issues concerning web search, like those related, for instance, to the so-called above mentioned "filter bubble", were not considered in this experience, which was basically meant as a preliminary activity aimed at enhancing the students' awareness of limits in their web search performance, and - by contrast – of the huge potential of well-devised refined queries.

Before describing the methodology and discussing the results, it is indeed important to stress once again that one of the goals of the present study, as discussed especially in section 1, was to show that ICT fluency is often more assumed than tested, with specific reference, in this case, to web search skills. The activities were accordingly designed so as to increase the students' awareness of the fact that a superficial, "satisficing", approach to web search might lead to incorrect assumptions about the search results. On the contrary, by carefully refining their queries through inductive reasoning, the students could achieve the best results, and – by doing so – have a rewarding experience in data-driven learning.

Participants

The school experience reported in this section involved a group of 15-17-year-old EFL students (12 male – 10 female) from a scientific high school in Bari, in the South of Italy[1]. All students had been studying English as a foreign language at school for 5-7 years. This does not count English

[1] The school involved is Liceo Scientifico "G. Salvemini" (Bari, Italy). The author wishes to express her gratitude to Prof. Francesca Palumbo for her invaluable support and to all the students who took part in the project.

language learning in primary school as this tends to be highly varied in Italy in terms of goals and standards and does not take into account any private language courses the students may have individually attended. Their L2 level was determined to be between A2-B1 on average, as is typical in secondary schools in Italy. None of them had had previous experience with corpora or other forms of data-driven learning.

Procedure

Since it is not always easy to arrange a visit to pre-tertiary schools, even for research purposes (see Crosthwaite's introduction to this volume), an English language teacher (Prof. Francesca Palumbo) acted as a 'mediator' and 'supervisor', as students were invited to perform specific tasks through a written assignment which I had designed for them (see Appendix). The set of activities proposed to the group of students was introduced with a note disclosing (very briefly and explained in simple terms) basic key concepts concerning quantitative evidence in empirical approaches to the study of languages. Using everyday examples (e.g. which is more common between "heavy rain" and "thick rain"?) the students were invited to consider frequency of occurrence on the web as a parameter to make an informed decision when in doubt about alternative wordings for similar concepts. Subsequently, the students filled in an online feedback questionnaire on the activities.

The methodology adopted was basically an inductive one. Rather than giving the students instructions about the potential of web search for linguistic purposes according to the principles discussed in previous sections of this chapter, the students were faced with problems to be solved and were only later introduced to very basic web search tips from a corpus linguistics perspective. The aim was indeed not so much to instruct them in the first place, but to reveal shortcomings in their expedient approach to web search and arouse interest in the potential of query refinement. Indeed, when asked whether they considered themselves well-equipped in terms of web search strategy, and whether they considered themselves capable of using web search to find solutions to language questions, they all seemed to be very confident, as exemplified by the following sample for responses to Question 2 of the survey:

> ***Would you be able to use web search to find a solution to problems concerning language use?***
> *Yes, I would because I always use online dictionaries*
> *Yes, I would because I often use the Internet to study.*
> *Yes, I would because on the Internet there are a lot of examples*

Yes, I would because it's so easy

Answers of this kind suggest that the group of students felt completely at ease with the task at hand and that introducing tasks based on web search would not be problematic, but also – significantly - that they were probably expecting nothing new from this experience. It is against this background that the classroom activities were designed, in the first place, as a way to 'defamiliarize' web search, by showing how much is taken for granted – which, in fact, needs a more reflexive approach.

Warming up: Calibrating the instrument

In its most basic form, using the web as a source of evidence of attested usage can entail a simple search for a single word. A patently simplistic use of the web for data-driven learning, searching for a single word is not devoid of interesting implications, and provides a good starting point for exploring the potential and the limitations of the web as a source of quantitative and qualitative data with students. It is in fact of crucial importance for anyone wishing to use the web as a linguistic resource to be aware of the relative importance of quantitative data elicited from the web. A case in point is spell-checking, which can be used as one of the most appropriate examples to show how misleading web search can be. Considering that younger people tend to see the web as a new *ipse dixit* whose reliability is not called into question, the warming up phase in our virtual teaching unit was devoted to what has been termed 'calibrating the instrument' (e.g. Figure 1):

1) Warming up

In this experience we are going to use the web as a tool for "measuring" language use. As is the case with many other tools for measuring things, we need first of all to 'calibrate' the measuring instrument, to see whether and to what extent it is accurate. More specifically, we are going to test the web for accuracy when providing quantitative evidence of attested usage.

- Search the web for a number of commonly mis-spelt words (e.g. accomodation, beautifull, independant, unforseen...). How many hits have you found for each of them? How many hits for the correct spelling? What are the implications of this simple experience, as to the reliability of the web as a linguistic resource?

Fill in a table with your results:

QUERY Wrong	# of hits	QUERY Correct	# of hits
accomodation		accommodation	
beautifull		beautiful	
independant		independent	
unforseen		unforeseen	

Figure 1 – Section "Warming up" from the worksheet (See Appendix)

When faced with the results of a web search for misspelt words and their correct counterparts, which inevitably laid bare the extent to which the web is also a repository for incorrect information (one only needs to consider the impact of such results as 86100000 vs 1650000000 hits for *independant* vs *independent*), the students were shocked. How could the web be used to find reliable answers to language questions, when there was so much 'noise' in it? This is what they were led to discover through the following activities.

Task 1. Evidence of attested usage for collocations

Being aware of 'noise' in the web is a good starting point for students to be engaged in DDL using the web as corpus. This will give them a clue about the importance of devising a more refined and complex query to obtain the best results. Furthermore, it will make them more cautious in handling quantitative data as they become aware of how much that is patently 'wrong' can be found online. Starting from these premises, the first activity was focused on collocation, and was designed to introduce the students to the use of inverted commas to search for exact phrases. As discussed in literature on the web as corpus (Gatto 2009: 59-65) using the web to validate collocations can be a really rewarding experience, provided that the query is appropriate to the task. Now, again, rather than starting from explicit instructions, the students were invited to search the web on the basis of their supposed skills as digital natives. The question concerned the evaluation of "big", "heavy", and "strong", as collocates for "smoker". The students were invited to answer this question on the basis of intuition and compare their hypothesis with information retrieved from the web.

Nine students out of twenty-two rightly indicated "heavy smoker" as the most appropriate collocation on the basis of their previous knowledge. However, when asked to support their answer on the basis of quantitative evidence form the web, they were surprised, as the most frequent collocation seemed to be "big smoker". Indeed, none of the students were aware of the possibility of using inverted commas to search the web for exact phrases, so they were faced with irrelevant results from the point of view of collocation, with "big smoker" and "strong smoker" finding more matches than the target collocation "heavy smoker".

Which adjective among “big”, “heavy”, and “strong” would you consider most appropriate in the collocation with “smoker”?

☐ BIG SMOKER ☐ HEAVY SMOKER X☐ STRONG SMOKER

Now use the web to support your answer:

QUERY	# of hits
big smoker	132.000.000
heavy smoker	38.200.000
strong smoker	68.500.000

TIP: use inverted commas to search for exact phrases

QUERY	# of hits
“big smoker”	64.900
“heavy smoker”	10.600.000
“strong smoker”	15.400

Figure 2. Section “Evidence of attested usage for collocation” from the worksheet (completed by one student)

As already argued, the methodology favoured an inductive approach. Not introducing advanced web search tips before the task, but only later, was therefore functional to the goals of the study. Firstly, it demonstrated how naïve the approach to web search is by so-called digital natives; secondly it stimulated in the students an interest for advanced web search tips through practice, rather than on a purely theoretical basis. It goes without saying that experiences like this could be the gateway for a discussion of more complex issues concerning collocations, which younger learners would probably better understand after having had the possibility of appreciating it, so to speak, ‘hands-on’.

Task 2. Testing translation candidates

Building on this task, and taking “heavy rain” vs “thick rain” as translation candidates for the Italian “pioggia fitta” (literally “thick rain”) as a new example, it was easy for our students to compare the frequency of occurrence of the two alternatives on the web *(see Exercise 3 in the Appendix)*. The students noticed that “thick rain” was fairly frequent on the web, with 89,800 hits, though less frequent than “heavy rain”. The problem was then how to interpret these results. Should the not

negligible number of matches for “thick rain” be considered nonetheless as clear evidence of attested usage, or did it instead provide support to claims about the total unreliability of web frequency data as a source of linguistic information (especially if the hits for “thick rain” are compared to the 34,100,000 hits for “heavy rain”)? It was hard to provide an answer to a question like this, since language itself constantly changes, and what is untypical today might become the standard of tomorrow.

Furthermore, this example opened the door to more consideration concerning the potential unreliability of quantitative data elicited from the web on a different level. In fact, one student observed that “Heavy Rain” is the name of a popular videogame, so the number of hits for this phrase was certainly inflated for this reason. This provided an opportunity to test the role of the minus sign to discard undesired results, and to design a new query: *“heavy rain” -videogame*. Reference to the videogame also suggested that the results retrieved for a query like *“heavy rain”* could not necessarily contain only pages from the UK and in English, which paved the way for an introduction to advanced web search parameters such as language and provenance of the query. Eventually the more complex query submitted through the advanced search interface for the exact phrase “heavy rain” without the word “videogame”, retrieving results only in English and from the UK, resulted in only 2,750,000 results, which still outnumbered the results for “thick rain”.

This task was used as a gateway to proper data-driven learning using traditional corpora. Faced with uncertainties about the results obtained from the web, the students appreciated the opportunity to query a more reliable resource for evidence of attested usage. Following a simple but detailed description on how to access the BNC through Mark Davies’ Corpora (see Follow up section in the Appendix) the students were asked to search the BNC for the two phrases and see which one is more frequent. The results provided by the BNC seemed to be indisputable. Over 200 occurrences for “heavy rain” and only 2 for “thick rain”, 1 of which was part of the phrase “thick rain forest”. The difference between the web as an anarchic though useful resource, and a real corpus, had been made clear.

Task 3. Query refinement in CLIL activities

Having introduced the students to some basic strategies for using the web as a source of evidence of attested usage, thus improving their search skills, attention shifted to the importance of producing more complex queries so as to enhance the relevance and reliability of the results retrieved (as

discussed in Gatto 2014: 69-70). This was done with reference to CLIL activities where Italian students are asked to use English as a vehicular language for other subjects (e.g. Physics).

A case in point was the search for the most appropriate wording for the concept of *momentum* in physics, as reported in the following task:

In a paper by a schoolmate you read the following sentence:

*Momentum, the **motion quantity** of a body, is given by the product of its mass and velocity.*

Do you think that "motion quantity" is appropriate here? Can you use web search to support your opinion?

Consider alternative wordings. Can you use the web to find out which of the following, if any, is more appropriate in physics?

- motion quantity
- quantity of motion
- movement quantity
- quantity of movement

Take note of your queries and of the results you find (add lines if you need):

QUERY	N of hits	Notes

Figure 2: CLIC activity on alternative wordings for the same concept (see Exercise 4 in the Appendix)

In this case the students were already familiar with exact phrase match and all of them searched the web accordingly. The results seemed to suggest that the most frequent phrases were "quantity of movement" and "quantity of motion". However, doubt was instilled in the students' mind as to the relevance of these results to the domain of physics, leading to the second part of this task:

Was it easy to find an answer on the basis of the number of hits for your query? Are you satisfied with you results?

If you haven't done so, repeat this exercise putting each phrase in inverted commas ("motion quantity", "quantity of motion", and so on) and adding the word *physics* to your search. This will increase the relevance of the results you get to the domain of physics, by including in the results only pages which also mention phisics, which is what you are interested in:

e.g. "motion quantity" physics

Does the number of hits you get change?

Finally you can restrict your query to "Google Scholar" or "Google Books", to increase the reliability of the results. Report your data in the following table and comment.

QUERY (Google Books)	N of hits	Notes
"motion quantity" physics		
"movement quantity" physics		
"quantity of motion" physics		
"quantity of movement" physics		

Do these results validate or invalidate previous results? Comment.

Figure 3: Exercise on query refinement (see Exercise 4 in the Appendix)

Again, the results were daunting, and the students were enthusiastic about the impact that refining their query had had on the relevance and reliability of the results. In particular, a search restricted to Google Books forced them to consider other criteria besides frequency alone. Going back to the initial task, i.e. the validation on the phrase "motion quantity", the fact that it seemed to be attested and validated in published books posed a new question: could "motion quantity" be still considered as a useful variant for the most frequent form "quantity of motion"? At this stage, new qualitative criteria entered into play, as the students realized that most of the results for the query *"motion quantity"* could not be considered as reliable evidence of attested usage simply because they did not refer to the phrase "motion quantity", but were merely the result of spatial proximity of the words, often separated by a comma or some other punctuation mark, and not part of the same phrase (see Fig. 4 below).

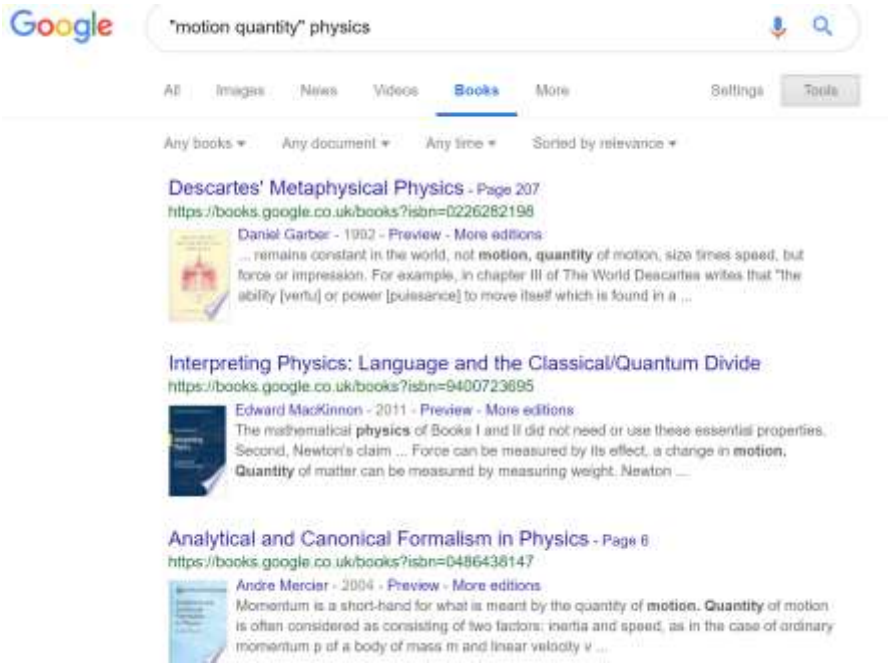

Figure 4: Search for "motion quantity" in Google Books

Conclusion

This chapter started with a reflection on the nature of younger learners as 'digital natives', and eventually welcomes the revised version of this same concept by Prensky himself in terms of 'digital wisdom' as a better way to address the relationship between younger learners and technology, which plays a key role in DDL. Digital wisdom is not just a matter of being clever in manipulating technology, but rather entails "making wiser decisions because one is enhanced by technology". This means, in turn, one may rely on intuition, but let intuition be "informed, inspired, and supported by digital enhancements and by the additional data digital tools provide" (Prensky 2009: 26). The reported examples in this chapter suggest that using the web as a corpus for DDL can positively contribute to this process. By learning how to use advanced web search, students can query the web in a more effective way in order to support intuition with data, so as to find data-driven solutions to problems concerning language usage. Furthermore, using the web as a corpus forces them into a process of progressive query refinement where each operator in an advanced web search can be reinterpreted from a corpus linguistics perspective. This can, in the long term, trigger an interest in corpus linguistics and pave the way towards a competent use of proper corpus query tools, thus representing a potential gateway to all other forms of DDL.

References

Baeza-Yates R. & B. Ribeiro-Neto (2011). *Modern Information Retrieval*. Addison-Wesley

Baroni, M. & S. Bernardini (2006). *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit

Battelle, J. (2005). *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. London: Nicholas Brealey Publishing

Boulton, A. (2015). Applying data-driven learning to the web. In A. Leńko-Szymańska & A. Boulton (eds), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins, p. 267-29

Buckingham, D. (2011), "Technology, Education, and the Discourse of the Digital Native: Between Evangelists and Dissenters", in M. Thomas, IX-XI

Chinnery, M.G. (2008). 'You've Got some GALL: Google-Assisted Language Learning'. Language Learning & Technology, **Vol.12, 1, 3-11**

Deschryver M, & R. Spiro, "New forms of deep learning on the web", in Zheng, Cognitive effects of multimedia learning, Information Science Reference-IGI Global, NY

Eu J. (2017), Patterns of Google use in language reference and learning: a user survey, in *Journal of Computers in Education*, 4(4): 419-439

Gatto, M. (2009), *From Body to Web. An Introduction to the Web as Corpus*. Roma – Bari: Editori Laterza University Press *on-line*

Gatto, M. (2011), 'The body and the web. The web as corpus ten years on', *ICAME JOURNAL*, 35: 35-58

Gatto, M. (2014), The web as corpus. Theory and practice. London: Bloomsbury

Geluso, J. (2011). 'Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing'. *Computer Assisted Language Learning*, 26, 2. 144-157

Han, S., & Shin, J.-A. (2017). Teaching Google search techniques in an L2 academic writing context. *Language Learning & Technology*, 21(3), 172–194

Kennedy, G. E. & Judd, T.S. "Beyond Google and the "Satisficing" Searching of Digital Natives", in M. Thomas (2011), 119-135

Kilgarriff, A. (2001), 'Web as corpus'. in *Proceedings of the Corpus Linguistics Conference (CL 2001)*, University Centre for Computer Research on Language Technical Paper, Vol. 13, Special Issue, Lancaster University, 342-344.

Kilgarriff, A. and Grefenstette, G. (2003). 'Introduction to the Special Issue on the Web as Corpus', in *Computational Linguistics*, 29, 3, 333-347

Levene, M. (2010), *An introduction to search engines and web navigation*. John Wiley & Sons

Lüdeling, A., Evert, S. and Baroni, M. (2007), 'Using Web data for linguistic purposes', in M. Hundt et al. (eds.), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 7-24

Maniez, F. (2007) 'Using the Web and computer corpora as language resources for the translation of complex noun phrases in medical research articles', in *Panace@: Revista de Medicina, Lenguaje y Traducción*,9, 26, 162-167

McEnery, T. & Wilson, A. (2001), *Corpus Linguistics*, Edinburgh: Edinburgh University Press

Parser, E. (2011). *The Filter Bubble. How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Group US.

Prensky, M. (2001a). Digital natives, digital immigrants. *On the Horizon*, *9*(5), 1–6.

Prensky, M. (2001b). Digital natives, digital immigrants, Part II: Do they really think differently? On the Horizon, 9(6), 1–6.

Prensky, M. (2009). Homo sapiens digital: From digital immigrants and digital natives to digital wisdom. Innovate, 5(3).

Prensky, M. (2011), "Digital Wisdom and Homo Sapiens Digital", in M. Thomas, 25-35
Sha, G. (2010). 'Using Google as a super corpus to drive written language learning: a comparison with the British National Corpus'. *Computer Assisted Language Learning*, 23(5), 377-393.
Shei, C. C. (2008). 'Discovering the hidden treasure on the Internet: using Google to uncover the veil of phraseology'. *Computer Assisted Language Learning*, 21(1), 67-85
Sŏrgo, A. et al (2016). Attributes of digital natives as predictors of information literacy in higher education: Digital natives and information literacy, *British Journal of Educational Technology* 48(3)
Tapscott, D. (2009) *Grown up digital: How the net generation is changing your world*. New York: McGraw-Hill.
Thomas, M. (2011). *Deconstructing digital natives*. New York: Routledge
Thomson, P. (2013). The digital natives as learners: Technology use patterns and approaches to learning, *Computers & Education*, Volume 65, July, 12-33
Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: Benjamin
Wu, S., Franken, M., & Witten, I. (2009), 'Refining the use of the Web (and Web search) as a language teaching and learning resource', *International Journal of Computer Assisted Language Learning,22(3), 247-265.*