Full reference: Pérez-Llantada, C. (2024). Identity construction in digital communication for public engagement in science. Discourse Studies, 0(0). https://doi.org/10.1177/14614456241255267

Identity construction in digital communication for public engagement in science

Abstract

This article explores identity construction in citizen science web texts. Keyness and concordance analyses show that these texts reflect, construct and negotiate identity construction in various ways to ultimately support citizen participation in scientific processes. Scientists primarily construct a professional identity through self-representation markers ('we' pronouns). They present themselves as credible professionals, aligned with the socially established values and ideologies of the scientific community. In addition, they appear to create a collective identity to promote empathy and to encourage and maintain citizen engagement in scientific processes. However, processes of indexicality and relationality – that is positioning and dialogism – reveal that the construction of a professional identity is consistently made more relevant than a collective expert-crowd identity, thus exposing overt power asymmetries. Even when dialogue between experts and nonexperts is encouraged, all the monologic and dialogic interactions established by the hypertext structure of the project sites reflect power imbalances.

Keywords

Digital genres, discourse analysis, discourse and interaction, open science, positioning, power asymmetries

Introduction

Linguistic studies of digital genres in professional science communication have recognised that identity construction is a strategic discursive tactic to achieve the intended rhetorical goals of these genres. In investigating identity in academic homepages, Hyland (2018) argues that when describing their academic trajectories, scientists construct an institutional identity aligned with the values, ideology and practices of their scholarly community. Yang's (2017) study of audioslide presentations associated with online journal articles claims that scientists construct an authoritative identity, expressed by exclusive 'we' pronouns (explicitly excluding the addressee(s)), activity verbs and evaluative markers, to emphasise the value of their research findings. Similarly, in a study of online data articles (Pérez-Llantada, 2022) argues that scientists project their stance on the value of their research data using an assertive style, mainly realised by 'we' pronouns and epistemic expressions, specifically deontic modals.

With regard to digital genres of public science communication, studies on identity are somewhat less conclusive. Some suggest that identity construction serves primarily to engage audiences. Kirkup (2010) states that in academic blogging, scientists create a collective identity using a subjective style that seeks proximity and consensus with the audience. This identity is used to build connections with the scientists' disciplinary community and special interest sub-communities. According to Scotto di Carlo (2014), TED Talk presenters use 'we' pronouns to show their personal commitment to the topic of their talk and by this means negotiate solidarity with their public. Zhou and Hyland (2019) compare research articles and blogs, discussing the use of attitude markers and self-mentions to establish an egalitarian relationship between scientists and their audiences. However, some studies have argued that constructing an authoritative identity is fundamental in these digital genres. Mauranen (2013) explains that scientists discussing scientific controversies in science blogs often use interpersonal markers such as boosters and evaluative language resources, as well as self-representation markers (‘we’ pronouns) in order to express certainty in their arguments, which creates a sense of moral authority. Luzón (2018) argues that research group blogs use ‘we’ pronouns to showcase researchers’ competence, promote their work, and align with their disciplinary community. In their analysis of science blogs, Zhou and Hyland (2020) also conclude that claiming authority through identity construction is essential when discussing scientific issues.

Studies on digital genres that specifically aim to engage the public in science, either through donations or through direct participation in scientific processes alongside the scientists, have provided evidence that identity construction is a complex discursive strategy serving multiple functions. It is widely accepted that interpersonal strategies are effective discursive resources for constructing credible identities and building trust in scientific research. This has been argued in the case of the construction of “real-life superheroes” in online medical fundraising campaigns (Paulus & Roberts, 2018) and the construction of credible (authoritative) personas in online science crowdfunding projects (Mehlenbacher, 2017). Other studies have made a similar point, such as those by Pérez-Llantada (2023) and Reid and Anson (2019). They argue that in citizen science communication, scientists intentionally downplay their scientific expertise to make scientific content more accessible to non-expert and lay audiences, bridging the knowledge gap between experts and citizens. Orpin’s (2019) study concludes that in posts (tweets) that recontextualise scientific reports in a citizen science project, scientists use assertive expressions of epistemic stance (possibility and certainty), self-mentions, modal verbs and attitudinal markers to convey credibility and trust.

The above studies confirm that identity construction serves important rhetorical goals in digital science communication aimed at public communication of science. However, the generalisability of the studies is somewhat problematic because they do not fully explain the discursive processes of identity construction that underlie online social interaction aimed at a broader understanding of scientific issues. Only a few implicitly point in this direction. Examining the use of inclusive ‘we’ pronouns (i.e., those including the addressees) Scotto di Carlo (2014) notes that this specific identity construction redresses power asymmetries between scientific experts and lay audiences. Reid and Anson (2019: 220) suggest that in participatory science activities, the scientists acknowledge the social demand for knowledge and educate the public by providing them with the scientific knowledge they lack. ‘Dissemination’, defined as transfer of scientific information, and ‘education’ of publics without expert knowledge of scientific issues are in fact the first and second rungs of the “science communication ladder”, and are considered low-level participatory science communication (Gascoigne et al., 2022: 20). Both represent the first two rungs of the citizen science communication ladder, in which citizens participate in collecting and analysing data, but not in interpreting data or co-producing scientific knowledge. This limited participation results in citizens having little power in the process. Examining how identity is constructed in the discourse and the extent to which it establishes lay-expert hierarchies and boundaries in digital genres of public science communication can provide important insights into understanding the “exercise of citizenship” (Picardi & Regina, 2008: 3). Today’s socio-historical context increasingly advocates the “democratisation of science” (Bartling & Friesike, 2014; Kimura & Kinchy, 2016).

The social interactionist approach to identity online

Full reference: Pérez-Llantada, C. (2024). Identity construction in digital communication for public engagement in science. Discourse Studies, 0(0). https://doi.org/10.1177/14614456241255267

Although linguistic approaches to identity have proved insightful in identifying how language reflects identity construction through resources of self-representation, stance and evaluation, social interactionist approaches to identity (e.g., Bamberg et al., 2011; De Fina, 2011, 2019) delve deeper into the complexity of identity construction, seeing identity as constructed through dialogical processes that involve interaction with others, whether implicit or explicit. This approach to identity construction thus requires in-depth analysis of the discursive processes of indexicality, local occasioning and relationality that point to elements of the social context (De Fina, 2011, 2019). Identity's indexicality processes are traced at the level of discourse semantics, specifically through features of style and lexis (i.e., specific words and expressions). Processes of local occasioning determine how identities are constructed and negotiated in a given context of interaction and how the actors involved in such interaction recognise each other and negotiate their identities. As De Fina (2011: 271) puts it, indexicality and local occasioning processes disclose "how identities are enacted and communicated through linguistic behaviour in contextualised ways and how people go about understanding and negotiating them." This author also explains that identity can be created and negotiated in processes of relationality (i.e., in utterances conveying authorial positioning and dialogism). Relational processes can in turn reveal multivocality, or multiple "voices representing different identities" (De Fina, 2011: 273), which supports the view of discourse as polyphonic. De Fina uses Goffman's concept of "footing" in narrative analysis to explain multivocality, specifically the roles of the speaker based on whether they are i) the 'animator' (the person physically producing the utterance), ii) the 'author' (originator of the utterance), or iii) the 'principal' (the person positioning towards the content of the utterance). Importantly, De Fina (2019: 5) further argues that identity processes in digital environments complicate the identification of identity construction due to the many possibilities of reconfiguring different voices on the hypertext.

This article applies the social interactionist approach to identity construction in order to complement the linguistic perspective and better understand how identities —including social, personal, collective, and situational or locally occasioned ones— are created and negotiated in digital genres of public engagement in science. In line with De Fina (2019), I argue that tracking identity types in digital genres of science is a complicated endeavour. As Askehave and Nielsen (2005) suggest, the hypertext structure of digital genres offers various rhetorical possibilities to meet the demands of the genre. Finneman (2016: 1-2) also makes a related conceptual argument, claiming that the production of digital materials results in a variety of knowledge formats with "particular configurations of hypertext features". Building on this, I aim to demonstrate that online citizen science projects have a unique configuration in the digital medium, integrating diverse contexts of interaction that reflect different ways of constructing identity. In online citizen science projects, the design of the hypertext includes a predetermined layout that establishes a hierarchy of content (i.e., menu pages and subpages). As in other web texts, the content on the pages/subpages is arranged in text blocks (Askehave & Nielsen, 2015: 125). The pages feature various multimodal texts (e.g., images, charts and diagrams, photographs and embedded videos). Meaning making is self-contained within each page/subpage, but in all pages/subpages content is semantically linked to the content of the other pages/subpages. In addition, one of these pages integrates an online discussion board for the exchange of comments and feedback.

In examining health discourse, Schryer and Spoel (2005: 253) maintain that genres involve complex forms of agency and identity, which makes it worthwhile to examine issues of power in relation to citizen science discourse, given its social impact. This study applies the linguistic and social interactionist approaches to identity to analyse the discursive construction of identity in exemplars of citizen science projects online. The hypertextual properties of the digital medium are hypothesised to be central to understanding these web texts and crucial for addressing identity construction in this digital genre. The following research questions guided the investigation:

RQ1. What is the semantic salience of self-representation markers in online citizen science projects?
RQ2. What are the recurring semantic associations of self-representation markers?
RQ3. How do these markers relate to discursive processes of indexicality, local occasioning and relationality (positioning and dialogism) for identity construction?

Methodology

Online citizen science projects are becoming increasingly popular as a means of engaging the public in scientific research. These projects rely on electronic platforms that provide scientists with tools to communicate with the public. These platforms have rather similar hypertext designs and integrate interactive tools so that citizens can carry out specific tasks online (e.g., image identification and classification) and interact with the scientists. This study analyses the websites of citizen science projects from Zooniverse, which is currently the largest citizen science portal with over 2.7 million registered volunteers. Each project site includes a homepage that features a text block containing the scientists' request for citizens' help (Tag Line). At the bottom of this page, the Researcher Quote text briefly outlines the social impact of the project. To the right of this text, the About [name of the project] text summarises the project's aim and impact. The About page of the homepage menu gives access to information about the research context, the team, the project results, educational applications, and frequently asked questions about how to participate in the project. The Classify page of the menu features three embedded texts: Tutorial, Task, and Field Guide. Task is a short text (e.g., *Are there trees in this image?*) that triggers the citizens' classification of images using metatext buttons (e.g., *Yes*/*No*/*I don't know*). The Tutorial provides a preview of the workflow to guide citizens through the classification process. To support citizens, each Field Guide consists of entries describing the types of subjects that are involved in the classification process. The Talk page consists of an online discussion board where citizens post their comments and the scientists respond. The digital medium is thus the communication channel, also allowing for the combination of verbal texts and multimodal (visual, aural) texts such as images, photographs, tables and embedded multimedia texts. Supporting De Fina's (2011) claim that language is the most important semiotic system for expressing and negotiating identities, this study focused only on language and did not consider other semiotic modes that also contribute to identity construction. This limitation will be addressed in future research.

For this study the content of the 44 citizen science projects under the topic of "nature" available on *Zooniverse* was extracted and saved in .txt format to create an electronic corpus representative of citizen science discourse online. The corpus totalled 270,260 words. Considering the organisation of the hypertext, this content was further split into 4 sets of texts: Homepage, About, Classify, Talk. These were further divided into subsets of texts. The Homepage content (totalling 8,708 words) was subdivided into Tag Lines, Get Started texts, Researcher Quotes and About [title of the project] texts. The About page content (109,559 words) was subdivided into About Research, About the Team, About Results, About Education and About FAQ texts. The Classify page content

(117,025 words) was split into Tutorial and Field Guide texts. Finally, a total of 20 notes were extracted from the discussion board of each project's Talk page (34,968 words).

Tracing the discursive construction of identity was assisted by *AntConc* software tools (Anthony, 2019). First, keyness analysis was computed to pre-filter the data, identify the semantically salient (unusually frequent) linguistic forms in the corpus of citizen science projects in comparison with a reference corpus (the BROWN corpus) and then set initial hypotheses. This analysis was based on the default keyness values (log-likelihood (4-term) keyword statistic; $p>0.05$ (+ Bonferroni) keyword statistic threshold; dice coefficient to measure keyword effect size). Drawing on the existing literature on digital genres, it was initially hypothesised that, as in other digital genres (Hyland, 2005, 2011, 2018; Yang, 2017), 'we'-pronouns and related forms ('our', 'us') were key self-representation markers.

Based on the keyness results, *Antconc* frequency list tool was used to retrieve descriptive statistics, i.e., the absolute frequencies of language forms conveying identity in the four contexts of interaction established by the hypertext design of the projects (Homepage, About, Classify and Talk pages). Absolute frequencies (i.e., number of occurrences) were normalised per 1,000 words to calculate relative frequencies and then make results comparable across these pages. Coefficients of variation (CV) were calculated to identify dispersion, or the distribution of resources across the texts. A one-way ANOVA analysis of variance test (set at 95% confidence limits, $p<0.001$) was computed using Excel to calculate significance regarding the frequency of identity markers. Effect size measures (η^2 values) were also calculated to track significance across the different sets of texts. A Tukey's HSD post hoc multiple (pairwise) comparisons test was run to identify significance between group means.

To make the empirical findings more robust, a cluster analysis was subsequently conducted using Anthony's software. The purpose was to identify structure in scientists' language choices. First, 3-word clusters (i.e., especially frequent word combinations formed by words in sequence) of 'we' pronouns were searched for. A threshold of over 10 (>10) or 20 or 25 occurrences, a minimum frequency of 5 and a range of occurrence in at least 5 texts of all texts (i.e., the minimum number of files in which the cluster appears) were set to make the results more robust. While aiming at ensuring representativeness, low threshold parameters were set since the modular texts of the projects' pages/subpages were all relatively short. A 4-word cluster analysis was subsequently run to identify tighter associations of word combinations. Finally, the concordance lines of the 4-word clusters and the discourse features co-occurring in their immediate cotext were analysed to track identity indexicality, local occasioning and relationality processes. Specifically, formal and non-formal (colloquial) discourse features were considered signals of identity indexicality processes. The presence of discourse features of stance and evaluation was associated with relationality processes. The micro-social interactions held between scientists and citizens on the Talk page were considered instances of 'local talk', or processes of local occasioning.

Table 1 summarises the analytical procedures followed in this study.

Table 1. Corpus-assisted discourse analytical procedures

KEYNESS ANALYSIS	Identification of semantic salience of self-representation markers (overall and across contexts of interaction)
	→ keyword list (keyness values and positive/negative keyness)
	→ statistical analyses of the distribution of self-representation markers

CLUSTER ANALYSIS	Identification of recurring semantic associations of self-representation markers (‘we’-clusters)
	→ 3-word and 4-word cluster analysis (overall corpus)
	→ 4-word cluster analysis (across contexts of interaction)
DISCOURSE ANALYSIS	Identification of discursive processes of identity construction (overall and across contexts of interaction)
	→ analysis of concordance lines of 4-word clusters → analysis of discourse features in the cotext

The commentary on the findings took into account the fact that communication on the site was unidirectional, with the exception of the Talk page, which was bidirectional. The interpretation of the findings was also informed by the information processing circumstances established by the hypertext, as described in the aforementioned frameworks (Askehave & Nielsen, 2015; Finneman, 2016).

Results

Semantic salience of self-representation

Overall data showed that ‘we’ was the most frequent marker of self-representation when compared to the possessive adjective ‘our’ and the oblique pronoun ‘us’. The keyness analysis indicated that ‘we’ was unusually frequent in the citizen science projects in comparison with the reference corpus. Whereas the first person singular pronoun ‘I’, also a marker of identity, was unusually infrequent (747 occurrences and a negative keyness value of -438.53; d=0.0054), ‘we’ occurred 2,209 times, ranked 8th in the list of keywords and had a positive keyness value of +1338.08 (d=0.016). The possessive adjective ‘our’ scored 988 occurrences and had a keyness value of +610.35 (d=0.0072). The pronoun ‘us’, with 509 occurrences, ranked 94th (keyness value=+291.88; d=0.0037) (Table S1. Top 30 keywords with positive keyness). Table 2 also shows that ‘we’ was by far the most common form of self-representation in each of the contexts of interaction established by the hypertext configuration of the projects. The fact that ‘we’ is the first constituent of the sentence and controls the verb of the clause indicates that in all these contexts identity is constructed primarily through the doers of the actions.

Table 2. Distribution of self-representation markers

	WE		OUR		US	
	Absolute freq.	Comp. %	Absolute freq.	Comp. %	Absolute freq.	Comp. %
Homepage	130	47.97	77	28.41	64	23.62
About page	1197	59.73	525	26.20	282	14.07
Classify page	535	58.99	265	29.22	107	11.80
Talk page	347	72.29	98	20.42	35	7.29

The normalised frequencies revealed significant variation of self-representation markers across the different contexts of interaction of the project sites. ‘We’ was used much more frequently in the Talk texts (32.29 occurrences per 1,000 words) and in the Homepages (31.12) than in the About and Classify pages (19.22 and 7.76 respectively). As deduced from its low coefficient of variation (CV<1), a homogeneous use of ‘we’ in all the homepages suggests that this marker plays a prominent role in this context of interaction, presumably, the main entry point to the information of the project (Table S2. Coefficients of variation of self-representation markers). Furthermore, according to the one-way ANOVA test, the salience of 'we' was statistically significant across the different contexts of interaction (F=13.97; p-value=0.00). A very large size effect (η^2=0.1959) indicates a very strong relationship between the variable 'context of interaction' and this self-representation form (Table S3a. ANOVA results for 'we'). On the other hand, although the use of 'our' and 'us' was not homogenous in all the projects analysed (CV>1), there were significant differences when comparing the different contexts of interaction (F=15.39; p-value=0.00 and F=18.95, p-value=0.00 respectively; see Table S3b. ANOVA results for 'our' and Table S3c. ANOVA results for 'us'). Size effects were very high (η^2=0.2117 and η^2=0.2484 respectively). These findings are therefore of practical significance and were taken into account when interpreting the results of the subsequent analyses.

The distinct behaviour of the three markers ('we', 'our', 'us) in each context of interaction was further validated by the results of the post-hoc test. The use of 'we' varied significantly when comparing the Homepage texts and the About page texts, and the About page texts and the Classify and Talk texts (HSD=10.17; Q-alfa value=3.63; Mse=345.2289905; n=44) (Table S4a. Multiple comparisons for 'we'). There were also differences for the case of 'our' and 'us' (HSD= 4.35; Q-alfa value=3.63; Mse= 63.27761628; n=4 and HSD= 2.10; Q-alfa value=3.63; Mse=14.69476744; n=4 respectively) (Table S4b. Multiple comparisons for 'our' and Table S4c. Multiple comparisons for 'us'). If we assume that scientists use them to produce content in different situational contexts, the Zooniverse guidelines for project creation could explain this different behaviour. According to these guidelines, the homepage should feature a concise "a one-line call to action" (Tag line) and a brief introduction to generate interest in the project. The About page should provide information on the context of the project, the team, results, educational applications and FAQs. On the Classify page, the Tutorial should explain the workflow while the Field Guide should provide specific details to help citizens complete the classification process. The Talk page is intended for the exchange of comments and questions. Therefore, the guidelines may act as a metagenre that influences the linguistic choices of the scientists, including the choice of self-representation.

Semantic associations of 'we'-clusters

The cluster analysis showed that 'we', previously identified as the key marker of self-representation in the projects, computed twenty-one very tight relationships between 'we' and other words forming recurrent word combinations (Table S5. 3-word clusters of 'we', threshold level min. freq. >10, min. range >5), retrieving a total of 385 clustered patterns of language conveying self-representation. 'We' pronouns established tight relationships with necessity meanings (e.g. *we need your*; *we need to*), mental verbs (e.g. *we want* to; *we know that*), communication verbs (e.g. *we ask that*; *we are asking*) and modal verbs conveying certainty (*we will be*). It is also worth noting that in the top cluster (*we need your*) 'we' is related to the possessive form of the pronoun 'you', another keyword in the corpus (3rd in the keyness list) and also a marker of dialogism for audience engagement (see also Hyland, 2011; Luzón, 2018). The grammatical negator 'not' also formed an especially frequent 'we' cluster (*we do not*) (3rd in the rank) that introduced the main reasoning for the scientists' request for citizens' help (e.g. *there is a risk this information is lost if we do not identify these new gene* variants). 'We' also clustered with activity verbs in progressive form (e.g. *we are working*; *we are using*; *we are going*) and deontic modals (e.g. *we can use*). The modal 'can' was an unusually frequent keyword in the corpus (11th in the rank) and therefore also worth considering in the subsequent analysis (cf. Table S1). Even tighter word associations were found in the 4-word cluster analysis, which retrieved sixteen clusters occurring over 100 times (Table S6. 4-word clusters of 'we'). Half of these clusters were extensions of 3-word clusters (*we need your (help)*, *we would like (to)*, *we are looking (for)*, *we can use (the)*, *we need you (to)*, *we are able (to)*, *we ask that (you)*, and *we are going (to)*). These clusters established very tight relationships with 'you' and 'help', which ranked 3rd and 17th respectively in the keyness list (Table S1).

The cluster analysis showed that 'we' was associated with different semantic meanings depending on the context(s) of interaction in which the clusters occurred. The 'About' tag lines on the homepages used 'we' clusters expressing necessity (*need*), willingness (*trying*, *going*), and ongoing actions (*working on*, *working to*). On the About page, 'we'-clusters convey activity and capability meanings, while in the Tutorials and Field Guides, they are associated with direct requests (*ask*) and necessity meanings (*need*). The About Research texts and the Tutorial and Field Guide texts of the Classify page exhibited the higher presence of 4-word clusters, accounting for 40% and 37% of all 4-word clusters, respectively, and almost 80% of all 4-word cluster occurrences. This is an important finding since the systematic use of language in the About Research subpage (that informs about the context of the project) and the Classify page (where the citizens engage in data classification and are supported by the Field guide information) may not be coincidental. The high frequency of these clusters may be intended to aid citizens in processing and comprehending the information. In contrast, the Talk showed greater lexical diversity as indicated by the low frequency of 4-word clusters (Table S7. Distribution of 4-word clusters across contexts of interaction). This may be due to the fact that Talk is a dialogic (not monologic) discourse and its hypertext configuration serves interactive rather than informational and instructional purposes like the other sets of texts.

Processes of identity construction

Self-representation was expressed primarily through 'we'-clusters either conveying necessity meanings or expressing direct requests (e.g., *we need your help*, *we need you to*; *we ask that you*, *we ask you to*). These clusters reveal processes of relationality (dialogism) similar to those reported for other digital genres (Hyland, 2011 for the case of academic homepages; Luzón, 2018 for the case of research group blogs). In the Researcher Quote and About [title of the project] texts of the Homepage, as well as in the About research texts, these clusters construct an expert identity that indexes authoritativeness and shows alignment with the values of the scientific community. Example 1 illustrates how this identity is negotiated in the discourse by the explicit mention to their need for citizens' help ('your help') to improve 'our (the scientific community's) models'. Example 2 shows a similar linguistic behaviour: 'we', the research group, involved in a research activity, need citizens ('your help') so that the 'scientists' (the scientific community and the researchers involved in the project) better understand the phenomenon investigated. The immediate cotext of these requests further reveals how the researchers position themselves as experts in recognising the value of citizens' collaboration (e.g., *vital*; *will build*; *better understand*). The use of colloquial features making explicit the researchers' activity, along with the use of *to*-infinitive clauses clarifying the goal of citizens' collaboration also index identity negotiation. Exclamatives conveying emotion suggest that the expert-citizen relationship is negotiated. The accompanying cotext is semantically and pragmatically intentional as it mitigates the illocutionary force of these directive speech acts in order to generate interest, what the Zooniverse guidelines recommend.

(1) Scientists at Restor are using Artificial Intelligence (AI) to find trees and track the progress of forest restoration efforts worldwide. *We need your help* to train and improve our models! Repairing and restoring nature is vital for halting biodiversity loss and tackling climate change.

(2) We have collected hundreds of thousands of photos (and counting), and *we need your help* to identify all the animals in these photos! The identifications that you record will build a data set that scientists can use to better understand which animals exist in Gorongosa, where they are, how they behave, and how the ecosystem is responding to restoration actions.

The authoritative identity is pervasive in the other contexts of interaction, albeit nuanced in various subtle ways. For example, in the About [project title] texts of the Homepages, the clusters 'we are hoping to' and 'we are trying to', both of them formulated in progressive form, add immediacy to the research actions carried out by the scientists and tone down the assertive tone created by certainty modals (*will*, *be able to*), by this means constructing less assertive selves (examples 3-4). In the immediate co-text, statements of finality (*to*-infinitive clauses) explicitly state the goal of citizens' participation so that citizens understand the reasoning behind the scientists' request and engage actively. Expressions of appreciation (e.g., *thanks*) and emotional language (exclamatives) mitigate the authoritative tone conveyed by the exclusive we-pronoun.

(3) Thanks to the community science, *we are hoping to* get a first round of annotations for our collection of insect images collected in the Swiss Alps, which we will then be able to use to train a ML model.

(4) *We are trying to* understand how forests of kelp grow and change over time. We need your help to find these forests in pictures from space!

In the About (Research) page, there is a clearer delineation between the crowd (who contribute the data) and the experts (who use the data), which can be tracked by the use of the clusters 'we are able to', 'we will be able', 'we can use (the)' and 'we are going to'. As shown in examples 5-6, the immediate cotext of these clusters once again reveals discursive processes of relationality (positioning and dialogism). Evaluative markers (*new*; *extensive*; *really interesting*) overtly state that scientists can achieve their goals with the citizens' help serve to reaffirm an authoritative (expert) identity. Explicit reader mentions within expressions of gratitude (e.g., *thanks to your* […]) and a conversational discourse style (what-clauses, exclamatives) barely redress this authority positioning and the expert-crowd divide.

(5) Thanks to your contribution in the Deep Sea Explorers project, *we will be able* to perform some very new and extensive studies to try to answer some of these questions.

(6) What is really interesting is that, thanks to the acoustic receivers, we can detect the acoustic signals of marine mammals! Like bioluminescence, acoustic signals from mammals can be a noise in our detector. Thanks to your classifications, *we are*

going to better understand the specificities of this background and, in the end, our detector will be better tuned!

The expert-crowd divide is most apparent in the distinct situational identity of the Tutorial texts, which is crucial for encouraging citizens to participate. Example 7 shows how a welcome message is followed by a direct, assertive request to strategically direct citizens' attention to the Tutorial text. Yet, an indirect presupposition is made here. The scientists assume that, by accessing the Classify page, the citizens need clear instructions to participate in the classification. Explicitation of content is done through exemplification (in this case, visual support marked by the deictic *here*) and through the use of defining relative clauses to clariy technical terms. In this way, scientists provide clear and concise explanations of the workflow. Politeness markers (e.g., *please*) mitigate the pragmatic force of the directive (*draw*).

(7) In this workflow *we need you to* look at the images of lymph nodes, like the one shown here. Please draw around the germinal centres, which are light pink circles with a dark outline.

This compelling authoritative authorial voice is also found in the Tutorial and Field Guide texts, in which the clusters 'we ask that you' and 'we ask you to' reveal discursive processes of identity relationality and positionality. Examples 8-9 show how authorial positioning constructs an authoritative identity. The scientists portray themselves as experts guiding the crowd (*we ask you to*; *you will see*; *don't mark this*). They also take an overt stance on the complexity of the classification task (*can be difficult*), hence the high level of explicitness, clarification asides and detailed instructions found in the immediate cotext of these clusters. Emotional language (exclamatives) mitigates the evaluative positioning towards the content expressed in the utterances. In short, the scientists acknowledge a lack of knowledge, without attempting to establish a shared (collective) identity between experts and citizens.

(8) Welcome to Monkey Health Explorer! In this project *we ask you to* look at images of cells and mark the platelets (smaller dots) - see left of center on top of above image. You'll also see other things - including many red blood cells (round/donut-shaped), white blood cells (larger cells), and bits of cellular debris. Don't mark these!

(9) This task requires a sharp eye. It can be difficult to spot birds with brown, rocky backgrounds. *We ask that you* take your time and just do the best you can! Many other citizen scientists will review each photo, so that when we combine all the data we end up with accurate results! Try not to overthink your classifications - you always have your fellow citizen scientists to back you up!

This dominant authoritative identity appears to be negotiated in the Field Guide texts, where the scientists attempt to ascribe an expert identity to the citizens by acknowledging the challenges of the classification process. Yet, the recurring clusters 'we would like to', 'we are looking for' and 'we are interested in', point to indexicality and relational processes of identity construction. The utterances are very explicit and more factual (i.e., primarily message-oriented information) than interactional (i.e., using language to create solidarity with the citizens). Although the pragmatic force of the directives (e.g., *choose*) is toned down by abundant exclamatives clauses conveying emotion and the use of an informal language (e.g., conversational hedges such as *just*), the features of the co-text make power imbalances overt (example 10).

(10) *We are interested in* pulling out this specific type of interaction from these videos! Choose this category if you see an aggressive interaction between two animals - two woodpeckers, a woodpecker and a squirrel, or any other pairing. Sometimes it is a challenge to tell if an interaction is a fight or not, but just do your best!

Finally, attention to the 'local talk' on the Talk page, the context of interaction in which the scientists responded to the citizens' questions in encouraging and reassuring ways, uncovers processes of overt positioning and dialogism that reaffirm scientists' authoritativeness. As shown in examples 11-12, even when dialogue appears to be flowing smoothly, exclusive 'we' pronouns framed within assertive statements index an expert identity. The use of 'you' followed by activity verbs in the surrounding co-text of 'we'-clusters foregrounded the importance of citizen collaboration, suggests that there is an attempt to negotiate a collective identity. However, identity indexicality is signalled by the choice of a conversational tone and colloquial features such as informal greetings and vocatives, exclamatives, and expressions of gratitude (*thank you*; *thanks*), indicating that although the scientists use language to establish empathy with the citizens, the interactions still involve power asymmetries.

(11) Yes, this is correct. If you can infer what the month is supposed to be when they omit it, then *we ask that you* enter it. So, correcto! Thank you.

(12) Hi @NWiLab, the host number is MAPS8E1241 (the PIPR#, is the catalogue number). *We will be able* to recognize Taiwan, wherever you put it in either the country or Province fields. Thanks for helping on this project!

Discussion

The aim of this study was to investigate identity construction in citizen science web texts in order to better understand digital genres that aim to promote public engagement in scientific processes and a more democratic and participatory framework for science (Bartling & Friesike, 2014). Using corpus-assisted discourse analytical tools, the study has found that 'we' was the most prevalent linguistic marker of self-representation, far more prevalent than its related forms ('us'/'our'). Furthermore, by adopting a social interactionist approach to identity this study has provided evidence for the dialogic nature of identity, supporting previous claims in this regard (Bamberg et al., 2011; De Fina, 2011, 2019). The following sections discuss the main findings. How the digital medium creates specific contexts of interaction within the project sites, enabling the construction of particular situational identities, is also discussed.

Authorial choices for constructing online identities

The results of the study suggest that in the citizen science projects analysed the scientists create and negotiate identities through self-representation 'we'-pronouns. The genre literature contends that 'we' is a semantically and pragmatically salient keyword in digital genres of both professional and public science communication, as diverse as audioslides (Yang, 2017), data articles (author, 2022) academic homepages (Hyland, 2018) or microblogs (Orpin, 2019), to name a few. In this study, a socially-recognised professional identity was created through strong semantic associations (clusters) in segments of discourse in which the scientists recontextualised scientific content and showed alignment with the values and ideologies of the scientific community. The rich lexical variation of the 3- and 4-word cluster structures profiled the scientists as active individuals committed to their profession. This variation included necessity verbs, activity verbs in progressive form, verbs of thought, and deontic modals, all of which shaped scientists' authoritativeness. As illustrated above, self-mentions accompanied by epistemic (deontic) stances portrayed reliable professionals that build credibility and trust in scientific research. This socially ascribed identity was also evident given the presence of clusters such as 'we are hoping to', 'we are trying to', 'we would like to' that indexed scientists' interests in addressing problems and contributing to new scientific developments. This is a relevant identity type in other popular digital genres such as research group blogs (Luzón, 2018), science blogs (Mauranen, 2013) and science microblogging (Orpin, 2019; Reid, 2019). This identity was also identified by the concordance analysis and supports the view of identity as grounded in the social context that shapes the discourse (Author, 2023; Hyland, 2005).

This study has also found that in these projects there were some attempts to construct a collective identity through the use of 'we' in conjunction with 'you' pronouns, tentatively suggesting scientists' willingness to establish a sense of community. Like 'we', 'you' was an unusually frequent keyword according to the keyness analysis. Across all the web pages and subpages of the projects, 'we' was used most frequently in the Talk page interactions. According to the cluster analysis, 4-word clusters established very tight relationships with 'you' and 'help', both unusually frequent keywords in these web texts. Reader mentions accompanying 'we' pronouns in their immediate cotext are also key engagement strategies in academic blogging (Kirkup, 2010; Zhou & Hyland, 2019), research group blogs (Luzón, 2018) and TED talks (Scotto di Carlo, 2014). However, from the analysis of the immediate cotext of these clusters, it was clear that the rhetorical purpose of the projects was not to engage with the wider public by establishing an egalitarian relationship. Rather than building a sense of community by creating proximity and emotional connectedness, their authoritative tone, more or less modulated depending on each context of interaction, evokes the deficit model of science communication, as Mauranen (2013) also argues for the case of science blogs.

By applying the social interactionist framework, empirical evidence has been found to claim that the online citizen science projects instantiate what De Fina (2011: 273) refers to as multivocality, or polyphony of voices. The corpus data has shown that the scientists took the role of 'animators' in the Tag Lines of the projects homepages. The Talk interactions illustrated that they were physically engaged in the classification processes as the citizens do. Additionally, they performed the role of 'authors' or originators of the utterances, for instance when they described the context of their research, the team profile and the project results in the About page. In the Tutorial and Field Guide of the Classify page they provided detailed instructions. Furthermore, they positioned themselves towards the content and the value of scientific research in the Researcher Quote and About [name of the project] texts, thus performing as 'principals'. In the Talk page, the interactive page, they acted as both animators, authors and principals, revealing the complementarity of these authorial roles. Corpus evidence also strongly suggests that in addition to being engaged in the classification process and to physically posting comments and replying to citizens' posts, the scientists consistently expressed attitudinal stances. Multivocality is also a characteristic of other discourses online such as business discourse and health discourse (Askehave & Nielsen, 2005; Schryer & Spoel, 2005).

Slippers (2013) remarks that early-career scientists need to be trained in forms of communication that reach broad publics and promote socially responsible scientific research. The study findings suggest that scientists should be made aware of the different ways in which self-mentions ('we' pronouns) construct identities. Furthermore, they make it advisable to make scientists aware that professional identity has a strong cognitive component (De Fina, 2019) and can therefore be learned through academic enculturation within the disciplinary community.

Identity construction, friend or foe of public engagement?

By tracing processes of indexicality, local occasioning and relationality, this study has shown that identity construction appears to create power imbalances between scientists and citizens. As seen in the variety of interactions maintained in the web texts, both uni-directional and bi-directional, scientists took on the role of the scientific authority. Identity's indexicality processes, linguistically realised through a non-formal style and rich in interpersonal features, revealed the view of citizens as having no shared knowledge among the interactants, hence the reason why scientists sought to establish empathy and collegiality with citizens (see also Reid & Anson, 2019). The utterances containing the clusters 'we are able to' and 'we are interested in' conveyed positive values of the researchers, indicating that scientists viewed themselves as experts. The scientists foregrounded their capacity to accomplish the project goals and their credibility as professionals, which implicitly reflects power imbalances. In the Tag Line, Researcher Quote, and About page, the scientists' explanations of the rationales and motivations regarding the project goals were expressed in an assertive tone. Other especially frequent clusters (e.g., *we do not, we know that*) also portrayed assertive selves. Additionally, as deduced from the prevalence of the deontic modal 'can', which clustered with 'we', the scientists assumed that there is no expected confrontation (or counter-argumentation) and the citizens assume that the utterance is correct. Power imbalances also became evident when the scientists took on the roles of instructors and guides in the Classify page, where they employed direct requests, once again profiling themselves as experts. They did not acknowledge alternative positionings from citizens but rather positioned themselves towards the content expressed in their utterances. This non-deferential, unmodulated stance performed persuasively to initiate and maintain citizens' participation in the classification task.

The prevalence of epistemic stance over engagement also appears to be a rhetorical feature of citizen science projects online. This is not the case for genres of public understanding of science such as blogs (Kirkup, 2010; Mauranen, 2013), crowdfunding projects and medical campaigns online (Mehlenbacher, 2017; Paulus & Roberts, 2018), which use abundant interpersonal language resources. In contrast, in the citizen science texts analysed the voice of the scientific authority, portraying scientists as credible professionals, seems to be both a friend and a foe of public engagement in scientific processes. It builds credibility and trust but also implicitly establishes power asymmetries. Identity's indexicality and local occasioning processes might even tilt the balance towards the latter option, especially in the Talk page interactions. While the scientists recognised their collegiality with the citizens, indexicality and local occasioning processes suggest power imbalances in this particular context of interaction. The corpus evidence thus supports De Fina's (2019) contention that ideological positioning involves taking a personal and/or professional stance not only towards the propositional content of a given utterance but also towards its interactants, in this case the citizens.

As also observed in relational processes of identity, the construction of a collective identity was linked to the language of interpersonality. In the Tag lines, direct requests were formulated by 'we' clusters containing 'you'/'your'. In the About and Classify pages, self-mentions co-occurred with reader mentions in their immediate co-text. A collective identity was also co-constructed in the Talk interactions, where the choice of a colloquial tone created bonds with the citizens. These discursive strategies, that redress power imbalances, are also used in academic blogging (Zhou & Hyland, 2020), research group blogs (Luzón, 2018) and science blogs (Kirkpup, 2010), as well as in digital genres of professional communication such as audioslides and data articles (author, 2022; Yang, 2017). Other genres that construct a collective identity through interpersonal resources are citizen science projects and crowdfunding medical campaigns (Paulus & Roberts, 2018; Reid, 2019). Yet, it is worth recalling that in the citizen science projects analysed, the voice of 'scientific authority' was more prominent than interpersonal communication in the discourse. This voice served to build credibility and trust, while also recognising the unequal relationship between scientists and citizens. One might then conclude that, despite their scientific content and collegial tone, citizen science projects online stick to the first and second rungs of the science communication ladder, namely dissemination of scientific information and education on scientific issues (Gascoigne et al., 2022).

Situational identities and the affordances of the medium

In their comparative study of blog posts and journal articles, Zhou and Hyland (2019) contend that the rhetorical contexts afforded by the medium in the blogs account for language variation in these genre types. This study has provided insights into the way the hypertext design influences identity construction in online citizen science communication. Supporting the social interactionist perspective's view of identity construction as being context dependent (Bamberg et al., 2011), the study findings corroborate that the hypertext establishes distinct situationally-dependent identities. Zooming into processes of local occasioning —or local talk— allowed to identify how several contexts of interaction and situationally-dependent identities were strategically allocated in the different pages/subpages of the project sites. The statistical differences found with the ANOVA test and the Tukey post hoc test of multiple comparisons further support this claim. The Homepages represented one concrete communicative context that made relevant the collective 'we'/'you' identity. The succinct "call to action" Tag Lines, expressed by 'we'-clusters, overtly acknowledged the indispensable citizens' collaboration to accomplish the project goal. Yet, at the same time, the texts located at the bottom of the homepage, representing different contexts of interaction, depicted complementary identities. In the Researcher Quote and the more argumentatively elaborated About [title of the project] texts at the bottom of the Homepage, a socially-recognised identity was made relevant, that of the scientific community. As illustrated above, in these texts the scientists showed commitment to scientific inquiry and eagerness to tackle current scientific challenges. Askehave and Nielsen (2005) explain that web users' reception and perception of information depend on their entry point and preferred navigation paths. Assuming that the homepage is the main entry point to the projects, it is not surprising that the collective identity is made relevant in this page to accomplish the primary communicative purpose of the genre, to engage the public in scientific processes.

The study has also shown how the hypertext offers possibilities for reconfiguring identity in various ways. The situational context of the About page made relevant the professional (expert) identity, as this page served to depict the scientists as providers of information on the project's research context, goals and expected outcomes. This identity, particularly salient in the About Research texts according to the statistical analyses, gave credibility to the project and the research goals to be accomplished. In this subpage meaning making was built upon clusters embedding activity verbs in the progressive form, which portrayed scientists as active professionals, and hence credible individuals. On the other hand, locally occasioned roles —i.e., instructors and guides— in the Tutorial and Field Guide texts of the Classify page made relevant the expert identity in order to prompt and support citizens' participation. This very much coheres with the Zooniverse guidelines, which define the Tutorial as a place to explain to the citizen the workflow for classifying images and the Field Guide as the place to provide specific information to support citizens in the classification. This may explain why the authoritative voice was much more prominent in the Classify page than in the Tag Lines. Finally, the Talk page afforded the construction of a virtual space in which, apparently, there was no confrontation but rather alignment with the citizens. Talk was in fact the only page supporting bidirectional communication thanks to the web's interactivity affordances. In this virtual space, scientists acted as experts even if they engaged in a friendly dialogue with citizens. They provided simple confirmations of the citizens' classifications, politely corrected any incorrect classifications, and clarified any queries. In conclusion, the fixity of the hypertext design determined the way identities cohered and complemented each other. It also made them more or less relevant in each situational context to induce engagement behaviour in particular ways. The hypertext design therefore allows only a low level of citizen participation. This limited participation confirms that citizens have little power when it comes to collaborating in these online projects.

References

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]

Askehave, I., & Nielsen, A. E. (2005). Digital genres: A challenge to traditional genre theory. Information Technology & People, 18(2), 120–141.

Bamberg, M., De Fina, A., & Schiffrin, D. (2011). Discourse and identity construction. In S.J. Schwartz et al. (Eds.), Handbook of Identity Theory and Research (pp 177–199). Springer Science+Business Media. https://doi.org/10.1007/978-1-4419-7988-9_8

Bartling, S., & Friesike, S. (Eds.) (2014). Opening Science. Springer.

De Fina, A. (2011). Discourse and identity. In T.A. Van Dijk (Ed.), Discourse Studies: A Multidisciplinary Introduction (pp. 263–282). Sage.

De Fina, A. (2019). Discourse and identity. In C.A. Chapelle (Ed.), The Encyclopedia of Applied Linguistics (pp. 1–8). John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal0326.pub2

Finnemann, N.O. (2016). Hypertext configurations: Genres in networked digital media. JASIST, 1-10, online. https://doi.org/10.1002/asi.23709

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & Marchi, A. (Eds.), Corpus Approaches to Discourse: A Critical Review (pp. 225–258). Routledge.

Gascoigne, T., Metcalfe, J., & Riedlinger, M. (2022). The ladder of power: Science communication and citizen science. Revista Lusófona De Estudos Culturais, 9(2), 15–27. https://doi.org/10.21814/rlec.4059

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. Discourse Studies, 7(2), 173–192. https://doi.org/10.1177/1461445605050365

Hyland, K. (2011). The presentation of self in scholarly life: Identity and marginalization in academic homepages. English for Specific Purposes, 30(4), 286–297. https://doi.org/10.1016/j.esp.2011.04.004

Hyland, K. (2018). Narrative, identity and academic storytelling. ILCEA [Online], 31. http:///doi.org/10.4000/ilcea.4677

Kimura, A. H. & Kinchy, A. (2016). citizen science: Probing the virtues and contexts of participatory research. Engaging Science, Technology, and Society, 2, 331–361 https://doi.org/10.17351/ests2016.099

Kirkup, G. (2010). Academic blogging: Academic practice and academic identity. London Review of Education, 8(1), 75–84. https://doi.org/10.1080/14748460903557803

Luzón, M. J. (2018). Constructing academic identities online: Identity performance in research group blogs written by multilingual scholars. Journal of English for Academic Purposes, 33, 24–39. https://doi.org/10.1177/074108831772629

Mauranen, A. (2013). Hybridism, edutainment, and doubt: Science blogging finding its feet. Nordic Journal of English Studies, 13(1), 7–36. https://doi.org/10.35360/njes.274

Mehlenbacher, A.R. (2017). Crowdfunding science: Exigencies and strategies in an emerging genre of science communication. Technical Communication Quarterly, 26(2), 127–144. https://doi.org/10.1080/10572252.2017.1287361

Mehlenbacher, A.R. (2019). Science Communication Online. Engaging Experts and Publics on the Internet. The Ohio State University Press.

Paulus, T. M., & Roberts, K. R. (2018). Crowdfunding a "real-life superhero": The construction of worthy bodies in medical campaign narratives. Discourse, Context & Media, 21, 64–72. https://doi.org/10.1016/j.dcm.2017.09.008

Pérez-Llantada, C. (2022) Online data articles: The language of intersubjective stance in a rhetorical hybrid. Written Communication,39(3), 400-425. https://doi.org/10.1177/07410883221087486

Pérez-Llantada, C. (2023) 'Help us better understand our changing climate': Exploring the discourse of Citizen Science. Discourse & Communication, 0(0). https://doi.org/10.1177/17504813231158927

Picardi, I., & Regina, S. (2008). Science via podcast. JCOM 7(02), C05. https://doi.org/10.22323/2.07020305

Puschmann, C. (2014). (Micro)blogging science? Notes on potentials and constraints of new forms of scholarly communication. In S. Friesike & Bartling, S. (Eds.), Opening Science (pp. 89–106). Springer.

Reid, G. (2019). Compressing, expanding, and attending to scientific meaning: Writing the semiotic hybrid of science for professional and citizen scientists. Written Communication, 36(1), 68–98. https://doi.org/10.1177/0741088318809361

Schryer, C., & Spoel, P. (2005). Genre theory, health-care discourse, and professional identity formation. Journal of Business and Technical Communication, 19(3), 249–278. https://doi.org/10.1177/1050651905275625

Scotto di Carlo, G. (2014). The role of proximity in online popularizations: The case of TED talks. Discourse Studies, 16(5), 591–606. https://doi.org/10.1177/1461445614538565

Slippers, B. (2013). Public engagement should start early. Nature, 498, 299. https://doi.org/10.1038/498299d

Yang, W. (2017). Audioslide presentations as an appendant genre – Key words, personal pronouns, stance and engagement. ESP Today, 5(1), 24–45. https://doi.org/10.18485/esptoday.2017.5.1.2

Zou, H., & Hyland, K. (2019). Reworking research: Interactions in academic articles and blogs. Discourse Studies, 21(6), 713–733. https://doi.org/10.1177/14614456198669

Zou, H., & Hyland, K. (2020). “Think about how fascinating this is”: Engagement in academic blogs across disciplines. Journal of English for Academic Purposes, 43, 100809. https://doi.org/10.1016/j.jeap.2019.100809