

Fuzzy-Based Language Grounding of Geographical References: From Writers to Readers

Alejandro Ramos¹, Jose M. Alonso^{1,*}, Ehud Reiter², Kees van Deemter^{2,3}, Albert Gatt⁴

¹Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

²Department of Computing Science, University of Aberdeen, Aberdeen, United Kingdom

³Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

⁴Institute of Linguistics and Language Technology, University of Malta, Malta

ARTICLE INFO

Article History

Received 06 Jul 2019

Accepted 23 Aug 2019

Keywords

Natural language generation
Linguistic descriptions of data
Data-to-text
Geo-referenced data
Language grounding
Fuzzy sets

ABSTRACT

We describe an applied methodology to build fuzzy models of geographical expressions, which are meant to be used for natural language generation purposes. Our approach encompasses a language grounding task within the development of an actual data-to-text system for the generation of textual descriptions of live weather data. For this, we gathered data from meteorologists through a survey and built consistent fuzzy models that aggregate the interpersonal variations found among the experts. A subset of the models was utilized in an illustrative use case, where we generated linguistic descriptions of weather maps for specific geographical expressions. These were used in a task-based evaluation to determine how well potential readers are able to identify the geographical expressions grounded on the models.

© 2019 The Authors. Published by Atlantis Press S.A.R.L.

This is an open access article distributed under the CC BY-NC 4.0 license (<http://creativecommons.org/licenses/by-nc/4.0/>).

1. INTRODUCTION

The abundance of data that surrounds our daily lives has allowed the emergence of several disciplines focused on researching how to better communicate to people the relevant information held within raw data. One such field is natural language generation (NLG), which studies the problem of how to generate texts from data that can be useful to human readers [1,2]. Within NLG, systems that generate texts from non-linguistic data are known as data-to-text (D2T). This kind of systems have become rather popular in recent times, thanks to their extensive application commercially [2,3] in a wide variety of domains.

Most D2T systems provide texts or reports that describe time series data, and plenty of examples can be found in the literature, e.g., in weather forecasting [4–6], health [7,8], or industry [9], among many others (see the following reviews for more references [1,10]). Alongside time series, geographical data has also been treated in D2T, although not as extensively as time series data [11].

An essential task in the conception of D2T systems is language grounding, i.e., determining how words and expressions are anchored in data [12,13]. There are several ways for solving this task, such as using heuristics or machine learning algorithms on an available parallel corpora of text and data [14] to create models of the expressions of interest anchored on the underlying data [15], getting experts to provide these mappings, or gathering data from

writers or readers that can be used for the application of mapping algorithms.

For instance, D2T systems that generate texts from time series data generally include temporal expressions to refer to relevant events or patterns found in those data. Performing language grounding of temporal expressions in this kind of systems allows ensuring that generated texts include words whose meaning is aligned with what writers understand or what readers would expect [15,16]. Likewise, in systems that produce texts from geographical data, having a good geographical model is essential to generate expressions that refer properly to specific locations and regions of interest. Moreover, when outliers are present in any resource that is used to create the data-to-words mapping, building language models that remove inconsistencies can also be a challenging part of language grounding which can influence the development of a successful system [15].

Quite often, temporal and geographical expressions that need to be included in texts generated by D2T systems are vague, such as “in the evening” [4] or “southwestern areas” [17]. In situations where vagueness (and thus borderline cases and gradual concepts) is present, fuzzy sets have been proposed as a tool that allows to model linguistic expressions for NLG/D2T systems [2,18,19]. However, current existing D2T systems do not make use of such techniques, with the exception of GALiWeather [5], which provides a basic use of fuzzy sets to model temporal expressions and quantifiers.

*Corresponding author. Email: josemaria.alonso.moral@usc.es

In this context, this paper describes the methodology we have followed to perform a language grounding task of vague geographical expressions, which is part of a larger project to build a D2T system that generates descriptions of live weather data maps. Our approach involves a data gathering task, the creation of fuzzy models that aggregate the opinion of several experts, and a use case to check the appropriateness of the models for potential end-users of the D2T system to be developed.

It is common (and highly recommended) practice to perform language grounding based on expert data in order to build a D2T system. However, our proposal considers two additional, yet highly relevant, elements involved in human communication: vagueness in language and receivers of information, i.e., readers in our case. Thus, the novelty of our approach lies in the conjunction of these three elements: building *i) vague geographical expressions* based on *ii) expert data*, whose use is tested against *iii) potential readers* even before the actual D2T system has been developed.

The rest of this manuscript is organized as follows. Section 2 provides a critical review of the state of the art regarding the use of geographical references in D2T systems. Based on this review, Section 3 describes in detail the general context and the motivation of this work. Sections 4 and 5 explain the methodology we have followed to gather data from meteorologists and build fuzzy models for a set of geographical expressions. In Section 6 we present a practical use case and a task-based evaluation of these models with actual users. Finally, Section 7 provides some concluding remarks.

2. RELATED WORK

The field of NLG is extensive, with many different kinds of applications and systems for many different purposes like generation of textual reports from data, summarizing different textual sources, generating dialogue, narratives, or even poetry [1]. In our case, we are focusing on a very specific topic, D2T systems whose input data are geographically characterized and whose generated texts include geographical expressions that refer to the occurrence of certain events found in the data (e.g., “rain in Northern Spain” or “strong winds in Northeastern Scotland”).

Although this kind of references were introduced decades ago by the system FoG [6], RoadSafe can be considered the most recent and representative example of this kind of systems [20,21]. This

system used weather forecast data to generate textual forecasts for road maintenance purposes. These reports included temporal and geographical expressions that helped identify where and when certain relevant phenomena would take place, with the purpose of helping maintenance teams to keep the affected roads in a good condition (see examples of such expressions in Figure 1). Thus, developing RoadSafe required an extensive study on how to generate good geographical expressions that referred properly to the geography underlying the relevant information extracted from the input data.

Concretely, RoadSafe’s approach to generating geographical referring expressions is based on standard techniques from geographic information systems (GIS), that divide the underlying geography of the events using different spatial schemas, or *reference frames* [22], which in turn are composed of non-overlapping partitions (known as *descriptors*). For instance, the reference frame *Direction* is composed of the descriptors “northeast,” “southwest,” etc., and the frame *Coastal proximity* is composed of the descriptors “coast” and “inland.”

Once the numeric limits of each descriptor are defined for each frame using plate carrée coordinates (latitude and longitude), each data point can be characterized by a set of descriptors (e.g., “southwest” and “coast”). Then, an algorithm that generates geographical referring expressions selects the best descriptor set that describes the area formed by the subset of points that represent the event.

In more recent work, the geographical model used by RoadSafe was extended to include named spatial references, which, according to an study of different corpus collections, are the ones that truly predominate [11]. Moreover, the previous reference proposed a geographical referring expression generation algorithm that integrates the “name” reference frame with the rest of RoadSafe’s frames. For this, the proposal in [23] addresses the differences between absolute and relative references and a model based on mereology is proposed, where named descriptors are combined with descriptors from other reference frames (e.g., “southwestern Moray”).

The described approaches provide good models to generate proper geographical referring expressions based on a set of reference frames. However, in all cases the underlying models were developed based on a crisp partitioning of the geography of interest. Establishing exact limits between descriptors can be considered a non-intuitive assumption. For instance, according to the previous

Road surface temperatures will fall below zero **during the late evening and tonight** except in **areas below 100M**.

SW 10-25 gusts **this afternoon in southwestern areas**, veering WSW and increasing 15-35 after midnight, gusts 55-60 **during the evening and tonight** except in **areas above 500M**, increasing 20-45 then veering W **by early morning**, gusts 70-75 **tomorrow morning in most southern and central places**.

Wintry precipitation will affect most routes at first, falling as snow flurries **in some places above 300M** at first. Snow spreading **throughout the forecast period** to **all areas** and persisting **in some places above 300M** until end of period.

Figure 1 | Examples of texts generated by RoadSafe [17].

approaches, if a village is located in the border between two descriptors, depending on the spatial granularity of the data, it can be possible to assign opposing descriptors (e.g. “coast” and “inland”) to a pair of almost overlapping points. Thus, if we consider the vagueness inherent to even the simplest geographical references, such as “north,” we can not set borders of geographical references with precision [17].

Under this light, the limitations that characterize the geographical models used in previous D2T systems are an important incentive for searching other approaches that model imprecision or uncertainty when managing vague geographical references. In fact, the problem of vagueness treatment in geographical references is not limited to D2T systems, and a rather extensive discussion of this problem has existed in the GIS field since long ago [24]. In this sense, fuzzy sets theory has been applied in many cases to address vagueness in geographical concepts and spatial relations, e.g., [25–27].

Likewise, as we mentioned in Section 1, fuzzy sets have also been proposed in a more general way to be used in D2T systems to model vague terms [2,18,19]. To our knowledge, as of today the only D2T system which uses this kind of techniques and has been deployed in a real environment is GALiWeather [5]. There exist, however, an important number of use cases that apply fuzzy sets to extract linguistic descriptions of data, which in occasions are realized through template-based text generation methods [2,10].

3. MOTIVATION

Given the limitations in terms of vagueness modeling that characterize previous D2T systems that generate texts from geo-referenced data, the main motivation of this work is to improve the modeling of vague geographical references for NLG purposes. Our objective is to establish a methodology to create models of vague geographical references and algorithms of geographical referring expressions based on those models, through the use of fuzzy sets.

Since D2T is an eminently applied field, where advances at a research level are often related to the development of actual systems, in our case the methodology we propose is encompassed by the development of a D2T system to generate textual descriptions of the meteorological state in a map, using data provided by MeteoGalicia [28]. The generated descriptions will include geographical referring expressions that will allow to identify relevant meteorological phenomena in maps, such as temperature, wind, and sky state.

Within the development of the aforementioned D2T system, this paper describes the language modeling task of the geographical expressions of interest, which will be used in the text generation process. Moreover, the methodology here described is based on the ideas introduced in [29] and [30], which will allow to consolidate new ways of applying fuzzy sets theory to D2T in a practical way.

4. GATHERING DATA FROM EXPERTS

While it is common to perform language grounding from a parallel set of text and data, and analyze the meaning of words and expressions to be modeled, in our case this kind of resource was not available. On the one hand, we did not have access to an extensive data set (e.g., texts of weather forecasts and prediction data) on which to perform such analysis. On the other hand, our plan is to build a new D2T system to provide textual descriptions of live weather data. Thus, the approach we followed consisted in interacting directly

with the experts that manually produce written information about weather.

4.1. The Survey

Given that our aim is to model geographical expressions, we asked the head of the forecasting department to provide us with a list of the most frequently used geographical references for weather forecasts. Based on this list, we prepared a survey similar to the one described in [30], which was forwarded among the experts within the weather agency. In this survey, subjects were asked to draw on a map of the region of Galicia (displayed under a Mercator projection) a polygon representing a given geographical reference (see Figure 2).

Subjects were provided with a list of 24 descriptors, which appeared in random order. In this list, 20 out of the 24 descriptors are commonly used in the writing of weather forecasts by experts and include cardinal directions, proper names, and other kinds of references such as mountainous areas, parts of provinces, etc. (see Table 1 for a complete taxonomy of the descriptors). The remaining four expressions were added to study intersecting combinations of cardinal directions (e.g., exploring ways of combining “north” and “west” for obtaining a model that is similar to “northwest”), and are out of the scope of this paper, since for our current purpose we are only interested in descriptors that are actually used by meteorologists when they write forecasts.

4.2. Results and Qualitative Data Analysis

The survey was completed by eight experts,¹ resulting in a total of 192 drawings (160 without the non-considered descriptors). At a general level, we had hypothesized that experts would be rather consistent, given their professional training. We also expected some variation among the different answers and the reduced number of meteorologists participating in the survey. We have observed that this is clearly the case; the polygons drawn by the experts are rather concentrated and therefore there is a high agreement among them.

For instance, Figure 3 shows a representation of the answers given by the meteorologists for the cardinal direction “Eastern Galicia” and a contour map that illustrates the percentage of overlapping answers. Likewise, Figure 4 shows the same representations of the answers by experts for a different expression, “Atlantic regions.”

With the exception of cardinal directions, most of the descriptors in Table 1 focus on very specific areas, such as those categorized as “Mixed.” Some exceptions to this rule are the “Inland Galicia” descriptor, or the mixed coastal + name descriptors like “Atlantic Coast.” We have also found that, in a few cases, descriptors are almost equivalent to others. This occurs to “Atlantic Coast” and “Atlantic Regions,” but also to “Northern Galicia” and “Northern Third,” which are very similar in shape and size.

5. BUILDING FUZZY GEOGRAPHICAL DESCRIPTORS

In Figures 3 and 4, the contour plots could be taken as the basis for the semantics of their corresponding expressions, with a core region that is accepted by the majority, and a gradual decay as one

¹ These results were compiled into a data set resource, which has been thoroughly described [38] and can be retrieved online [39].



Figure 2 | Screenshot of the survey that the meteorologists completed (translated from Spanish).

Table 1 | List of geographical descriptors in the survey.

Category	Spanish	English Translation
Cardinal Directions	Norte de Galicia, Sur de Galicia, Oeste de Galicia, Este de Galicia, Tercio norte, Extremo norte, Noroeste de Galicia, Noreste de Galicia, Suroeste de Galicia, Sureste de Galicia	Northern Galicia, Southern Galicia, Western Galicia, Eastern Galicia, Northern third, Extreme North, Northwestern Galicia, Northeastern Galicia, Southwestern Galicia, Southeastern Galicia
Inland/Coastal Named Direction + Name	Interior de Galicia Rías Baixas, Comarcas atlánticas Oeste de A Coruña, Oeste de Ourense, Sur de Ourense, Sur de Lugo	Inland Galicia Rías Baixas, Atlantic Regions Western A Coruña, Western Ourense, Southern Ourense, Southern Lugo
Inland/Coastal + Name	Litoral Atlántico, Litoral Cantábrico, Litoral norte, Interior de Coruña, Interior de Pontevedra	Atlantic Coast, Cantabrian Coast, Northern Coast, Inland A Coruña, Inland Pontevedra
Mountain Area + Name	Áreas de montaña de Lugo, Áreas de montaña de Ourense	Mountainous Areas in Lugo, Mountainous Areas in Ourense

moves to the outer periphery of the outlined regions. Thus, in our case, fuzziness stems from the interpersonal differences among the meteorologists.

Following this notion, we have created fuzzy models that aggregate the opinions of the experts for each descriptor. The method we used for this modeling task is an improvement on the heuristic algorithm described in [30], which produced raw models based on point sampling and polygon intersection counting, without any enforced conditions. Our algorithm is preceded by a simple filtering of the drawings.

5.1. Filtering Outliers

As we discussed above, the drawings made by the meteorologists are very consistent visually, but in some cases we observed slight

inconsistencies. In order to maintain the high consistency that is found in the answers for most descriptors, we applied a simple filter to the polygon drawings, which consisted in discarding answers out of the $[\text{mean} \pm 2 * \text{standard deviation}]$ interval in terms of size and centroid location.

5.2. Characterization of a Fuzzy Geographical Descriptor

Using the filtered data, we aim to build fuzzy geographical descriptors which are simple and consistent.

Definition 1. Fuzzy geographical descriptor, G :

$$G = \{S, K, \mu_G\} \quad (1)$$

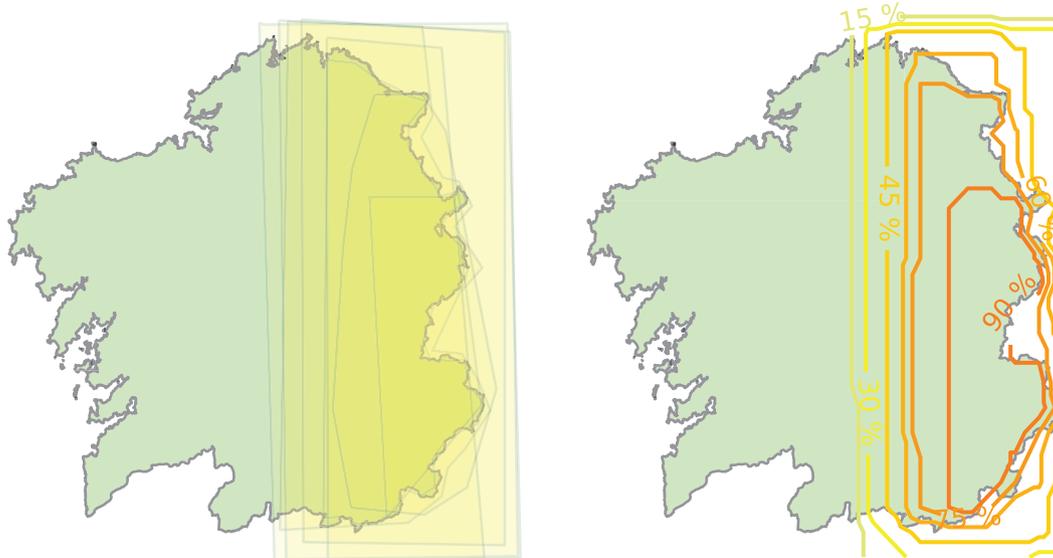


Figure 3 | Representation of polygon drawings by experts and associated contour plot showing the percentage of overlapping answers for “Eastern Galicia.”

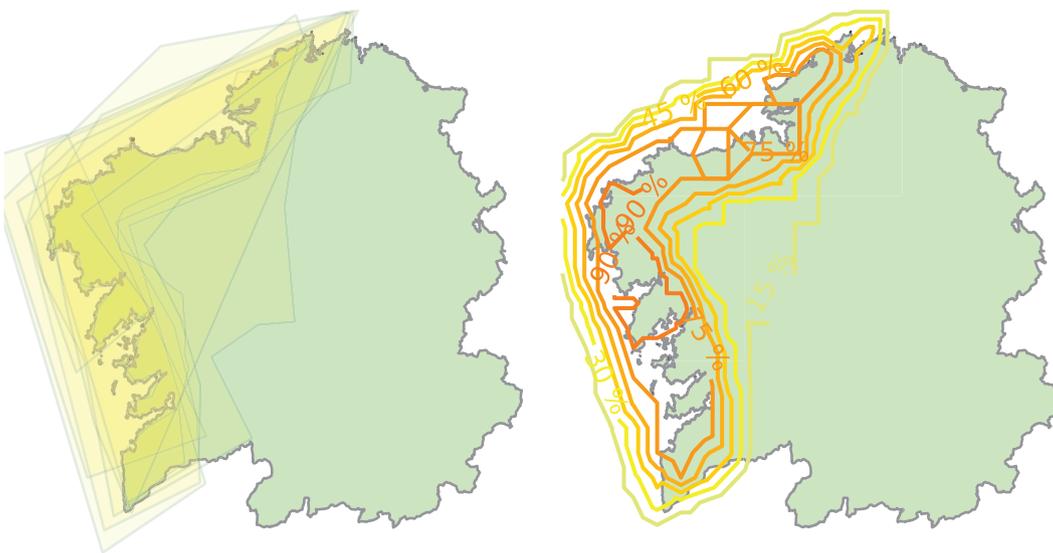


Figure 4 | Representation of polygon drawings by experts and associated contour plot showing the percentage of overlapping answers for “Atlantic regions.”

Formally, we define a fuzzy geographical descriptor G (e.g., “Southern Galicia”) as a set of three elements: a support area S , a kernel area K , and a fuzzy membership function μ_G , which evaluates the degree in which a point in a map ($p = (x, y) | x, y \in \mathbb{R}$) can be considered part of G :

$$\mu_G : \{\mathbb{R}, \mathbb{R}\} \rightarrow [0, 1] \tag{2}$$

Based on μ_G ,² K and S can be defined as:

$$K = \{p \mid \mu_G(p) = 1\} \tag{3}$$

$$S = \{p \mid \mu_G(p) > 0\} \tag{4}$$

Thus, K is the set of points (or region) whose membership degree is maximum with respect to G , while the support includes all points with a non-zero membership degree. However, in order to achieve consistent models, we need to apply the following restrictions to G :

$$\forall G \quad K \subseteq S \tag{5}$$

$$\forall \{p_i, p_j\} \mid p_i, p_j \in S, \quad p_i, p_j \notin K \text{ and } p_i \neq p_j \\ d(p_i, K) > d(p_j, K) \Leftrightarrow \mu_G(p_i) < \mu_G(p_j) \tag{6}$$

The previous conditions ensure that the fuzzy models are consistent, by avoiding the possibility of achieving disjoint K 's and S 's and

²For simplicity, we will refer to $\mu_G(p)$ instead of $\mu_G(x, y)$.

ensuring monotonicity for μ_G , where $d(p, K)$ is the euclidean distance to K from a point p in S .

5.3. Building Descriptors from Data

A fuzzy geographical descriptor G , like “Northern Galicia” is modeled based on the polygons drawn by the experts for that given expression. Formally, the collection of polygon drawings for a specific G is defined as:

$$R_G = \{P_1, P_2, \dots, P_n\} \quad (7)$$

Each P_i (with $1 \leq i \leq n$) represents a polygon, and n is the total number of polygons left after the initial filtering. Each polygon is composed of a set of vertices, defined under a plate carrée projection (longitude and latitude values).

The first task in order to model a given G is to determine its two constituents, i.e., S and K , as calculating both will allow us to characterize μ_G afterwards. For this, we define a grid of points $D = (p_1, \dots, p_i, \dots, p_{|D|})$, which is delimited by the geographical bounds of the underlying geography (the region of Galicia in our case). The distance between the grid points is determined by a parameter δ , which specifies a percentage of the region’s width and height. For instance, $\delta = 1$ means that the distance between a pair of grid points equals to 1% of the region’s height or width.

As it can be seen in the algorithm given below, in order to compute K and S for a descriptor G , we take D and the collection of expert drawings R_G , and we calculate the number of times each p in D is contained by a polygon P in R_G to determine the percentage of polygons that overlap in a given p . Based on the percentages computed for all points in D , we determine K using a simple majority voting approach. Thus, PK is composed of those points whose percentages are $> 50\%$, and PS covers all points where the percentage is $> 0\%$.

Require: D, R_G

Ensure: K, S

```

1: PCS ← ()
2: for all  $p_i \in D$  do
3: count ← 0
4: for all  $P_j \in R_G$  do
5: if  $p_i \in P_j$  then
6: count ← count + 1
7: end if
8: end for
9:  $pcs_i \leftarrow count/|R_G|$ 
10: PCS ← PCS  $\cup$   $pcs_i$ 
11: end for
12: PK ←  $\{p_i \in D | pcs_i > 0.5\}$ 
13: PS ←  $\{p_i \in D | pcs_i > 0\}$ 
14:  $K \leftarrow ConvexHull(PK)$ 
15:  $S \leftarrow ConvexHull(PS)$ 
16:  $bp \leftarrow argmax_{p_i \in S} (d(p_i, K))$ 
17:  $op \leftarrow argmin_{p_i \in D \text{ and } p_i \notin S} (d(p_i, b))$ 
18:  $od \leftarrow d(K, op)$  return  $K, S, od$ 

```

We chose to determine PK under the $> 50\%$ condition as we wanted to ensure that there exists at least a minimum majority that agrees

and that the modeling of the kernel is consistent across different descriptors. This was not the case, for instance, in [30], where the region of points with the highest overlapping would be selected, leading to kernels built on different agreement percentages. Furthermore, in our case the high consistency of the experts allowed us to maintain this condition, but in other situations with more divergent answers it would have been necessary to be more flexible or even to adopt a different modeling strategy.

Instead of just considering S and K as collections of points within D (what PK and PS are), we calculate their convex hulls, that is, the convex polygons that delimit their areas. This process removes any disjoint areas belonging to K that may emerge from divergences among the polygon drawings, as the convex hull aggregates all of them under the same area. Finally, we calculate od , which is the sum of the maximum distance from a vertex in S to K (bp) and the minimum distance from that vertex to another point outside S (op). This distance is used to define μ_G , together with K and S .

5.4. Evaluating a Fuzzy Geographical Descriptor

Based on the three elements returned by the Algorithm described in the Section 5.3, the function μ_G that evaluates a point p is defined as:

$$\mu_G(p) = \begin{cases} 1 & \text{if } p \in K \\ 1 - d(p, K) / od & \text{if } p \in S \text{ and } p \notin K \\ 0 & \text{if } p \notin S \end{cases} \quad (8)$$

The membership function μ_G is defined in a monotonous way. Under this definition, all points in S have membership degrees > 0 and this membership degree decreases as the evaluated point gets further located from K . Under these conditions, characterizing a fuzzy geographical descriptor G by determining its support and kernel allows us to provide simple models that are consistent and easy to interpret.

5.5. Examples

In order to illustrate the results of the modeling algorithm and explain the result of the modeling decisions that have been included in the Algorithm described in the Section 5.3 and in the definition of μ_G , we describe two different fuzzy models that result from aggregating the experts’ drawings for geographical expressions under different categories or reference frames in Table 1.

The first descriptor, shown in Figure 5, models the expression “Northern Galicia.” The kernel is slightly thicker toward the easternmost part of the region map, and does not seem to depend on the coast distance, but simply on latitude values. As a whole, the model can be considered straightforwardly regular, for both K and S . The simple majority condition also ensures a wider consensus on what the meaning of “north” is in this context, and the distance from K to the outer border of S accounts for 1/4 of the region extension. This provides a relevant area where points can be considered as part of “Northern Galicia” under different degrees. Similar models were

obtained for the rest of cardinal directions, with slight differences among them.

Other descriptors adopt different shapes, where S varies in width with respect to K . This is the case, for instance, of “Mountainous Areas in Ourense.” As Figure 6 shows, the shape of this descriptor is somewhat ellipsoidal and the distance between S and K is not constant, compared to Figure 5. This case also illustrates well how points in S which are close to its border have higher membership degrees than others that are further, due to their closeness to K .

It must be noted that the models shown in Figures 5 and 6 were built using $\delta = 1$, resulting in a grid composed of 10506 points that were used to determine K and S . Although in our case the number of answers to aggregate was reduced and increasing δ (to a certain extent) may not have had an important impact, in other contexts it can be feasible to change this value to create less precise models, depending on the granularity of the geographically localized data that we want to process (e.g., a 1 km weather forecast grid versus a 12 km grid).

6. USE CASE AND EVALUATION

As we explained in Section 3, the models here described are meant to be used to feed a geographical referring expression generation algorithm, as part of a D2T system to describe live weather data. Within this context, as a use case that shows the usefulness of the built models, we generated short textual weather descriptions for specific geographical descriptors. These descriptions were built based on the use of fuzzy linguistic descriptions of data [2,10,18] and the same kind of input data that will feed the D2T system.

Particularly, the input data in our setting is a collection composed of live observations for several weather variables (e.g., temperature, wind strength and direction, sky state), where each observation contains data for 314 municipalities (see Figure 7 for a visual example). The descriptions we generate focus on quantifying the sky state variable with respect to a geographical descriptor.

Then, using a set of representative cases, we performed a task-based user evaluation. Potential end-users of the D2T system were presented with maps similar to the one shown in Figure 7 and associated textual descriptions without explicit geographical references. With these materials, they had to choose the geographical descriptor that best suited both the description and the weather map. This allowed us to verify the appropriateness of a representative set of the modeled descriptors.

6.1. Generating Geo-Referenced Linguistic Descriptions

From a D2T perspective, fuzzy linguistic descriptions (or summaries) of data can be used as a tool for performing content determination tasks [2,5,10]. The content extracted from the raw data adopts the structure of protoforms composed of linguistic terms (modeled by means of fuzzy sets) [18], that still need to be properly verbalized.

In our particular case, we generated simple descriptions based on the use of type-II fuzzy quantified sentences [31], that relate two different variables (e.g., “most wet days were cold”) [10]. The specific kind of protoform we have used adopts the following structure:

$$Q \text{ } Xs \text{ in } G \text{ are } W \quad (9)$$

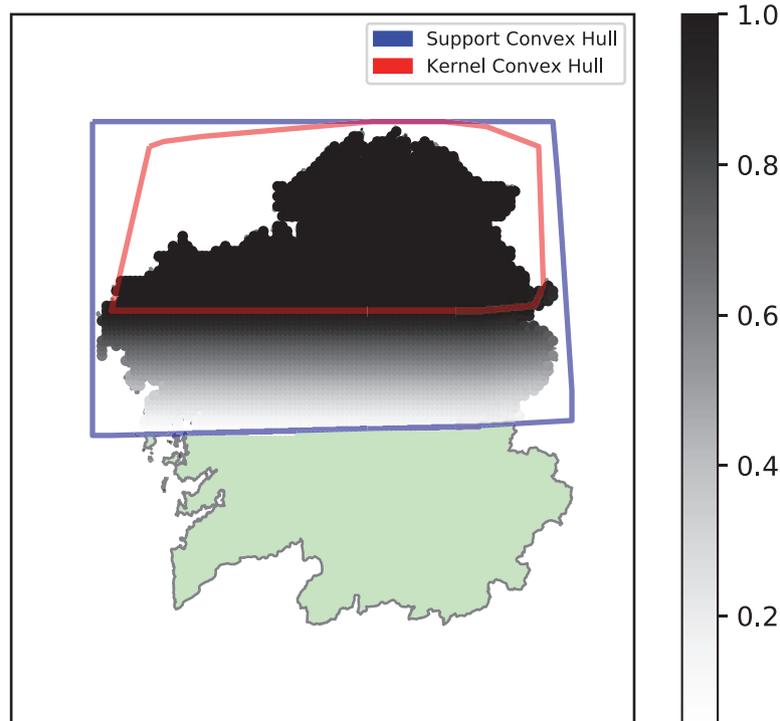


Figure 5 | Visualization of the fuzzy geographical descriptor “Northern Galicia.”

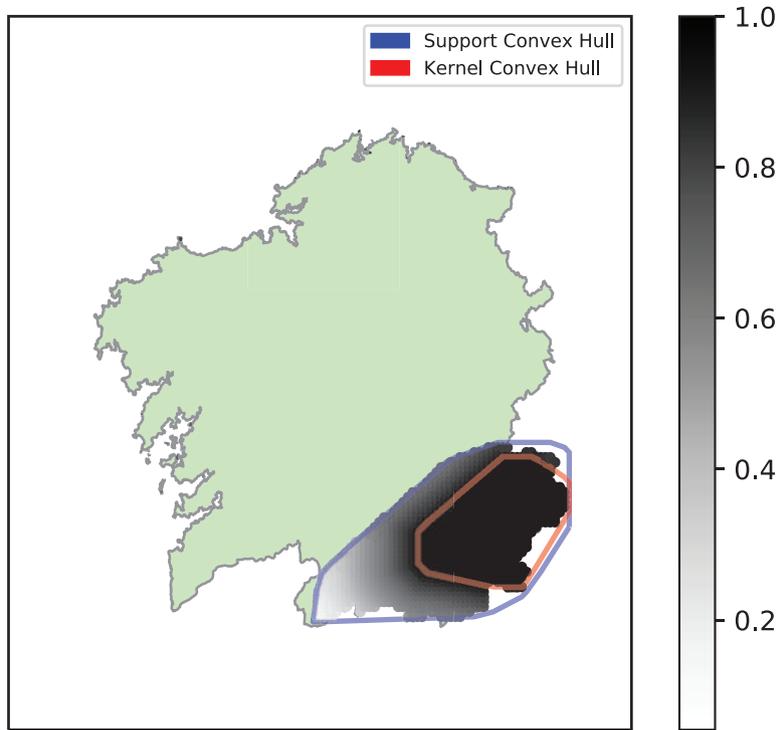


Figure 6 | Visualization of the fuzzy geographical descriptor “Mountainous Areas in Ourense.”

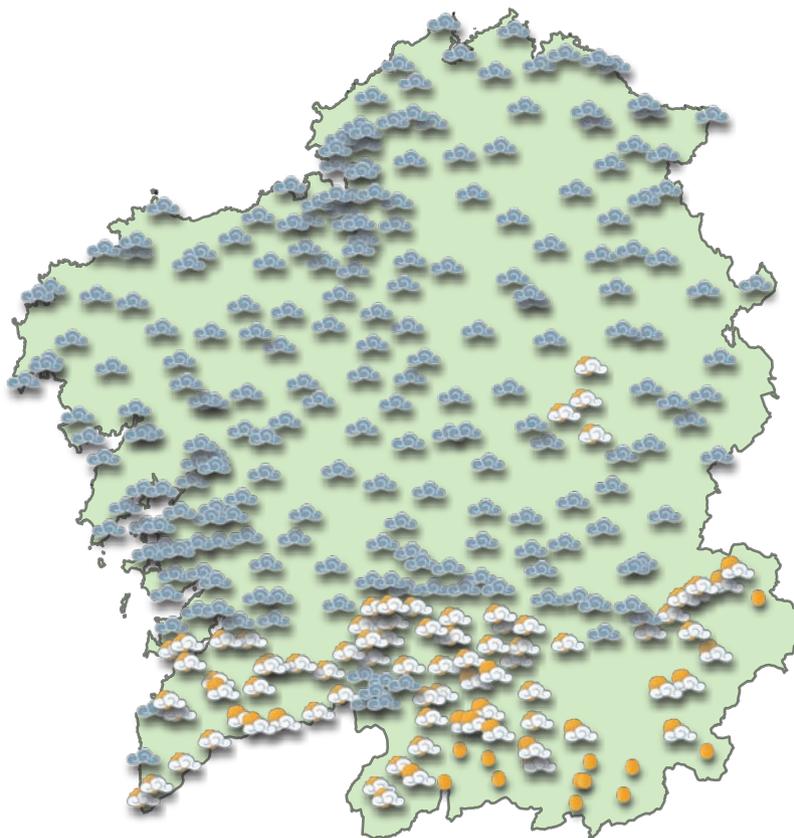


Figure 7 | Live weather snapshot (Galicia, 2017-06-02, 12:32 pm).

In this protoform, Q is a fuzzy quantifier, G is a fuzzy geographical descriptor, X is a referential set (the collection of data locations), and W is a sky state label, e.g., “most locations in Southern Galicia are overcast.” Sky state labels are crisply defined as numeric symbols (for instance, 101 means “clear sky,” and 111 means “overcast with rain”). In the case of the fuzzy quantifiers we used, we adapted the original crisp partition of three quantifiers used in RoadSafe [21] into the fuzzy quantifier partition shown in Figure 8.

In order to calculate the truth degree of sentences that follow the protoform defined in Eq. (9), a fuzzy quantification model must be applied [31]. We opted to use the quantification model GD [32,33], which was already tested in previous approaches [34,35]. This model fulfills several desirable properties. For instance, it is less susceptible to the problem of *aggregative behavior* [36], which is present in models based on sigma-count cardinalities; and has a time complexity of $O(n)$, which makes it more efficient than other *well-behaved* models like the F^A [37].

Under the GD method, the truth degree T of a geo-referenced protoform GRP is calculated as follows (see [33] for a more detailed definition of this method):

$$T(GRP) = \sum_{\alpha_i \in \Delta(W/G)} (\alpha_i - \alpha_{i+1}) Q \left(\frac{|(W \cap G)_{\alpha_i}|}{|G_{\alpha_i}|} \right) \quad (10)$$

For instance, in Figure 9 a description is given for the fuzzy descriptor “Extreme North” based on the weather map. The description is composed of several quantified sentences, which are characterized and ordered by the specificity of their quantifier (less specific ones are considered better in this case), their truth degree, and their relative cardinality. The sentences are individually realized using a template, which holds the textual structure of the protoforms, the actual linguistic expressions of the quantifiers and the sky state symbols.

6.2. A Task-Based Evaluation of the Fuzzy Descriptors

We evaluated the appropriateness of the fuzzy geographical descriptors by performing a task-based evaluation with potential users of

the D2T system we are developing. The main purpose of this evaluation was to check how well subjects were able to relate linguistic descriptions to the actual geographical references modeled by the descriptors, allowing us to see how they performed in clear and more ambiguous cases. Furthermore, this evaluation also provided a way to indirectly assess how well the geographical understanding of potential readers matches the models created from the experts’ opinions.

6.2.1. Materials

We selected a list of 5 geographical descriptors, making sure that they covered most of the whole region of interest and were representative of the different kinds of categories shown in Table 1. Specifically, we chose two cardinal directions (“Southern and Eastern Galicia”), one proper name (“Rías Baixas”), and two mixed descriptors (“Western A Coruña” and “Cantabrian Coast”). We also ensured that different degrees of overlapping existed among the descriptors with respect to the support of their fuzzy models, from none (e.g., between “Rías Baixas” and “Eastern Galicia”) to a full containment (e.g. “Cantabrian Coast” \subset “Eastern Galicia”).

For each descriptor we generated linguistic descriptions from a data set composed of more than 300 real live weather snapshots, which were collected between June and August 2017. We also generated maps for each snapshot, similar to the ones shown in Figures 7 and 9. The linguistic descriptions were generated following the approach described in Section 6.1, with some additional restrictions:

- Quantified sentences (GRP) were discarded when $T(GRP) < 0.5$ and $|GRP| = 0$. In case of a tie between sentences with contiguous quantifiers we chose the one whose quantifier represents a higher coverage.
- Within a linguistic description, sentences were ranked by the specificity of their quantifier (as described in Section 6.1), and then by $T(GRP) * |GRP|$ in a descending order, to help balance the overall influence of the truth degree and the cardinality when determining the suitability of the individual sentences.

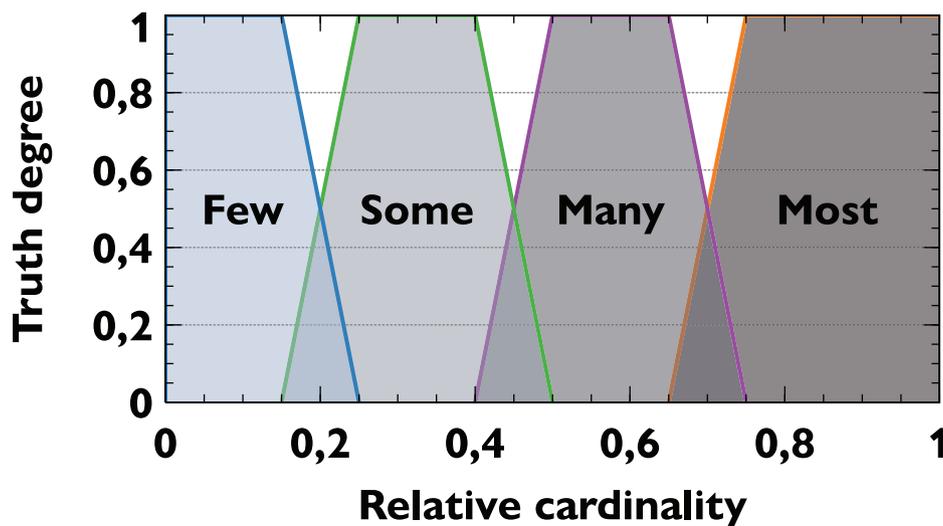
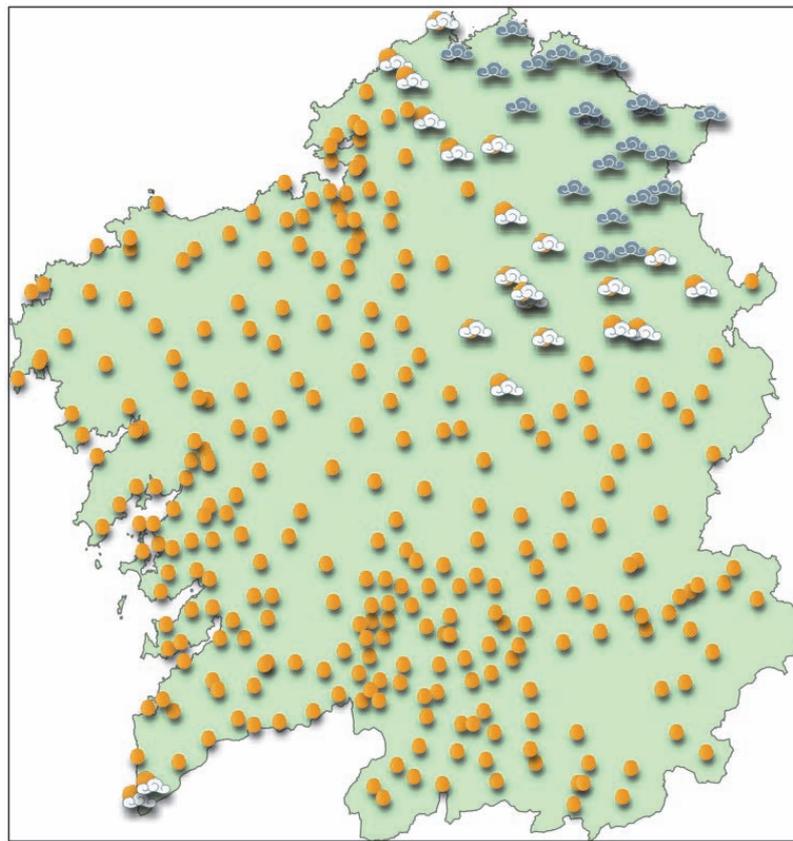


Figure 8 | Partition of fuzzy quantifiers used to compute the geo-referenced linguistic descriptions.



Sentence	T	Card.
Many locations in the extreme north are overcast	0.57	0.69
Some locations in the extreme north are partly cloudy	0.91	0.27
Few locations in the extreme north are clear	1	0.04

Figure 9 | Map example and associated linguistic description composed of several quantified sentences for the fuzzy descriptor “Extreme North” filtered and ordered by their truth degree (T) and cardinality (Card).

- Geographical references were replaced with generic expressions when we textually realized the quantified sentences (e.g., in Figure 9, *G* would be expressed as “target zone” instead of “extreme north”).

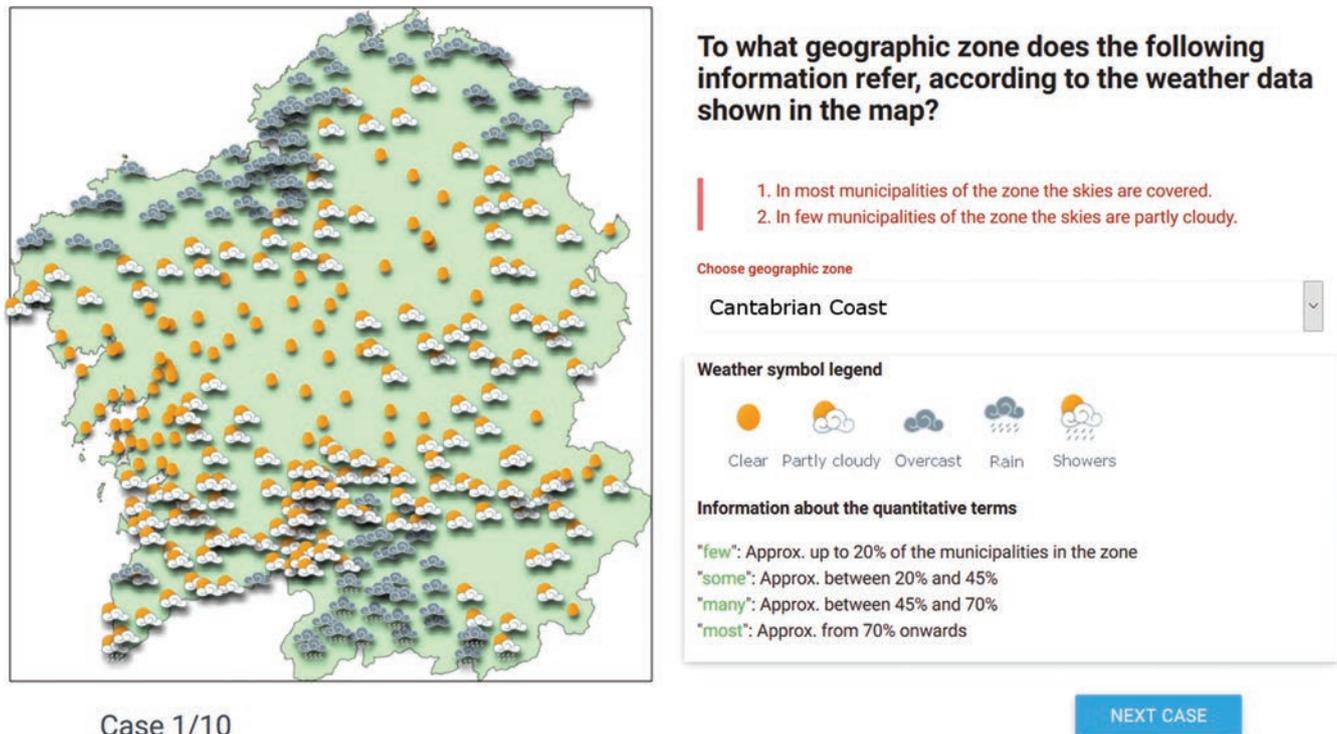
Based on this collection of parallel maps and linguistic descriptions, for each of the 5 descriptors we carefully selected one case where the matching between the description and the geographical descriptor could be clearly done according to the weather data shown on the map, and another case where this matching was ambiguous. For instance, in the latter case we looked for situations where the linguistic description could apply to other descriptors (e.g., with slight differences with respect to the quantifiers or the weather symbols that were mentioned). Thus, after this selection, the final set of evaluation cases for the survey comprised 10 maps and their associated descriptions (5 clear and 5 ambiguous cases).

6.2.2. Subjects

Our target was the group of potential users of the texts of the D2T system to be developed, namely Galician adults or anyone familiar with the geography of Galicia. These were recruited through university mailing lists and through family and relatives, ensuring that people from different gender, education backgrounds, and age ranges participated.

6.2.3. Procedure

Using the collection of weather cases and associated descriptions for specific descriptors, we created a web survey under a within-subjects design, where participants were given in random order a weather map and a linguistic description, and were asked to select in a drop-down list the geographical descriptor that best suited the



Case 1/10

Figure 10 | Screenshot of the web survey (translated from Spanish).

provided description, in accordance to the weather data shown on the map (see Figure 10). Given that the participants were native Spanish speakers, both the linguistic descriptions and the survey were created in Spanish.

As Figure 10 shows as well, we provided information about the weather symbols that appeared on the map, and a simplified textual definition of the quantifiers that were included in the sentences of the description, so that participants could focus just on the task of identifying the geographical descriptor the description referred to.

At the end of the survey, participants were asked to provide information about their gender, level of education, and age range. They could also write free-text comments about any aspect of the survey. Then, they were shown the percentage of cases they had guessed correctly, categorized into clear and ambiguous cases (they were not aware of this distinction).

6.2.4. Rationale, threats, and hypotheses

At first we had considered preparing a survey where we would provide maps and descriptions that explicitly referred to the geographical descriptors they were based on, asking users to rate how accurate the description reflected what was being showed on the map. However, this would have resulted in an important bias, as most users, knowing the geographical zone of interest in advance, would certainly have adapted their own mental models and agreed with the description.

Thus, instead we chose to have users make the effort of guessing the geographical references. While this is a more complicated task from the participants' perspective, it avoids having them focus on a specific zone before they can compare how well the description

To what geographic zone does the following information refer, according to the weather data shown in the map?

1. In most municipalities of the zone the skies are covered.
2. In few municipalities of the zone the skies are partly cloudy.

Choose geographic zone

Cantabrian Coast

Weather symbol legend



Information about the quantitative terms

- *few*: Approx. up to 20% of the municipalities in the zone
- *some*: Approx. between 20% and 45%
- *many*: Approx. between 45% and 70%
- *most*: Approx. from 70% onwards

NEXT CASE

matches the geographical reference. Also, our aim is not to get a rating of the descriptions themselves, since the final texts that the full D2T system will generate will be more elaborate in terms of both content and style. In this case, the descriptions are a tool to illustrate how the fuzzy geographical descriptors can be used to extract linguistic content and, in our evaluation, they allow users to identify the descriptors on a map.

Since we are focusing on specific content and the definitions of the geographical expressions rather than on full texts describing the weather state of the map (which will be the output of our D2T system), we expected some participants to find difficulties when interpreting the protoform-like sentences in the description. Also, the task of matching sky state labels and quantifiers with the data shown on the map was not trivial in several of the cases we chose, especially when a description was composed of several sentences. On top of this, we had assumed that participants would be familiar with the geographical references we chose, but their expertise might vary as well.

Considering these threats, we had hypothesized that *i*) subjects would perform well in clear cases, but not necessarily close to a proportion of 100% of correct answers; and *ii*) participants would perform better in clear cases than in ambiguous ones.

6.2.5. Results and discussion

The survey was completed by 40 subjects (400 answers in total, 200 for clear cases, and 200 for ambiguous ones). Participants guessed the geographical descriptors correctly in 155 out of the 200 clear cases, resulting in a proportion of 77.5%. As for the ambiguous cases, 83 out of the 200 answers matched the actual descriptor behind the description, resulting in a proportion of 41.5%.

We tested our first hypothesis (users perform well in clear cases) under a 1-sample proportions test. With a percentage of 77.5% correct answers, for values up to a 70% the null hypothesis was rejected, with a 95% confidence interval [0.71, 0.83] and a p-value of 0.025. Thus, we can state that participants performed well and guessed the right descriptors in a rather high proportion for clear cases. For our second hypothesis, we performed a 2-sample test for equality of proportions. In this case, the null hypothesis (the proportion of correct guesses being the same for both clear and ambiguous cases) was rejected with a 95% confidence interval [0.26, 0.45] and a p-value of $4.769 * 10^{-13}$.

Although we verified our two hypotheses, a more detailed analysis of the proportions for both situations is in order. In the case of the clear setting, for the total amount of clear cases that were answered, we can observe in Table 2 the ratio of descriptors answered by participants with respect to the correct answers. The most noticeable case where subjects failed to guess the correct descriptor was for “Eastern Galicia” (E.G., second column). Excluding the row which shows the correct guesses, all rows show for this case a proportion > 0 . This is especially relevant for R.B. (“Rías Baixas”). Although these descriptors do not overlap (see Figure 11), in this case they shared the same weather symbols, albeit in very different proportions. Thus, we believe this result might be due to some participants answering without paying excessive attention to the quantifiers in the description. Also, it is possible this case was slightly less clear than the rest.

Regarding the ambiguous cases, the lower rate of correct answers was to be expected, given that in these cases different descriptors (which in some cases overlapped and in others did not) shared the same weather symbols in more or less similar proportions. If we observe Table 3, in the case with the highest proportion of incorrect answers subjects confused “Western A Coruña” and “Eastern Galicia” (W.C., E.G.: 0.14). Although these descriptors do not overlap, the weather events reflected on the map were the same for both

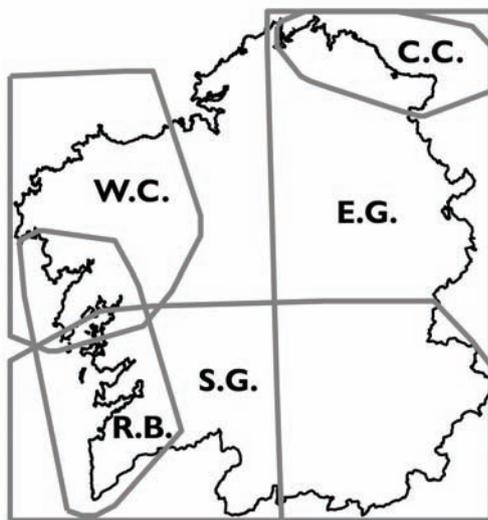


Figure 11 | Plot of the support areas for the geographical descriptors considered in the survey (S.G., Southern Galicia; E.G., Eastern Galicia; R.B., Rías Baixas; C.C., Cantabrian Coast; W.C., Western A Coruña).

Table 2 | Confusion matrix for clear cases, categorized by descriptors.

Actual \ Answer	S.G.	E.G.	R.B.	C.C.	W.C.
S.G.	0.18	0.02	0	0.01	0
E.G.	0.01	0.08	0	0	0.01
R.B.	0.01	0.07	0.18	0.02	0
C.C.	0	0.01	0.01	0.16	0.02
W.C.	0	0.02	0	0	0.16

C.C., Cantabrian Coast; E.G., Eastern Galicia; R.B., Rías Baixas; S.G., Southern Galicia; W.C., Western A Coruña.

Table 3 | Confusion matrix for ambiguous cases, categorized by descriptors.

Actual \ Answer	S.G.	E.G.	R.B.	C.C.	W.C.
S.G.	0.09	0.01	0.11	0.02	0.02
E.G.	0.01	0.02	0	0	0
R.B.	0.1	0	0.07	0	0.08
C.C.	0	0.04	0.02	0.14	0
W.C.	0	0.14	0	0.04	0.1

C.C., Cantabrian Coast; E.G., Eastern Galicia; R.B., Rías Baixas; S.G., Southern Galicia; W.C., Western A Coruña.

in rather similar proportions. Other interesting cases include subjects answering “Southern Galicia” for “Rías Baixas” (where R.B. is mostly contained in S.G.) or “Rías Baixas” for “Western A Coruña” (where R.B. and W.C. partly overlap).

Based on the obtained results, we believe the semantics of the geographical references we have modeled are appropriate for the D2T system we are currently developing. The difficulty of the task we proposed in our evaluation and the proportion of correct guesses in clear cases, which is close to 80% (and is statistically significant at 70%), are good indicators in this direction. Moreover, we expect that users will be able to easily identify weather events in the map in automatically generated texts, considering these will include the actual geographical references derived from the fuzzy geographical descriptors we created.

7. CONCLUSIONS

We have described in this paper an application of fuzzy sets to perform a language grounding task. Specifically, in the context of the development of a D2T system meant to generate geo-referenced weather descriptions, our work establishes a solid methodology for integrating fuzzy set-based techniques into the task of language grounding for D2T purposes.

Based on how expert writers understand certain expressions (geographical references in this case), we created simple and consistent fuzzy models that aggregate their perceptions. We have provided an illustrative use case where these models are used to generate linguistic descriptions. Based on this use case, we checked the models against potential readers of weather descriptions, evaluating how well readers relate to the references modeled on the experts’ preferences under different settings [16].

As future work, we plan to develop a geographical referring expression generation algorithm that is able to use the models here described. Based on the same kind of weather data shown in the

example figures in Section 6, such algorithm will generate expressions that combine different geographical descriptors to refer to different weather event areas distributed across a map. This algorithm will be the core of the new D2T system to describe live weather data.

CONFLICT OF INTEREST

There is no conflicts of interest.

Funding Statement

This research was also funded by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R and TIN2017-90773-REDT) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29 and “accreditation 2016-2019, ED431G/08”). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

ACKNOWLEDGMENTS

Jose M. Alonso is Ramón y Cajal Researcher (RYC-2016-19802) and member of the European Society for Fuzzy Logic and Technology (EUSFLAT). He is the corresponding author. He contributed jointly with the rest of co-authors to the conceptualization of the proposal, implementation, and evaluation.

REFERENCES

- [1] A. Gatt, E. Kraemer, Survey of the state of the art in natural language generation: core tasks, applications and evaluation, *J. Artif. Intell. Res.* 61 (2018), 65–170.
- [2] A. Ramos-Soto, A. Bugarín, S. Barro, On the role of linguistic descriptions of data in the building of natural language generation systems, *Fuzzy Sets Syst.* 285 (2016), 31–51.
- [3] Gartner, Neural Networks and Modern BI Platforms Will Evolve Data and Analytics, January 16, 2017. <http://www.gartner.com/smarterwithgartner/neural-networks-and-modern-bi-platforms-will-evolve-data-and-analytics/>.
- [4] S.G. Sripada, E. Reiter, I. Davy, SUMTIME-MOUSAM: configurable marine weather forecast generator. *Expert Update.* 6(3) (2003). <http://www.expertupdate.org/>
- [5] A. Ramos-Soto, A. Bugarín, S. Barro, J. Taboada, Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data, *IEEE Trans. Fuzzy Syst.* 23 (2015), 44–57.
- [6] E. Goldberg, N. Driedger, R. Kittredge, Using natural-language processing to produce weather forecasts, *IEEE Expert.* 9 (1994), 45–53.
- [7] J. Hunter, Y. Freer, A. Gatt, E. Reiter, S. Sripada, C. Sykes, Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse, *Artif. Intell. Med.* 56 (2012), 157–172.
- [8] A. Goldstein, Y. Shahar, An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data, *J. Biomed. Inform.* 61 (2016), 159–175.
- [9] J. Yu, E. Reiter, J. Hunter, S. Sripada, Sumtimeturbine: a knowledge-based system to communicate gas turbine time-series data, in: P. Chung, C. Hinde, M. Ali (Eds.), *Developments in Applied Artificial Intelligence, Lecture Notes in Computer Science*, vol. 2718, Springer, Berlin, Heidelberg, 2003, pp. 379–384.
- [10] N. Marín, D. Sánchez, On generating linguistic descriptions of time series, *Fuzzy Sets Syst.* 285 (2016), 6–30.
- [11] R. de Oliveira, Y. Sripada, E. Reiter, Designing an algorithm for generating named spatial references, in *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), Association for Computational Linguistics (ACL), Brighton, 2015*, pp. 127–135.
- [12] D. Roy, E. Reiter, Connecting language to the world, *Artif. Intell.* 167 (2005), 1–12.
- [13] E. Reiter, An architecture for data-to-text systems, *The 11th European Workshop on Natural Language Generation, Schloss Dagstuhl, Germany, 2007*, pp. 97–104.
- [14] J. Novikova, O. Dušek, V. Rieser, The E2E dataset: new challenges for end-to-end generation, in *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saarbrücken, 2017*. <https://arxiv.org/abs/1706.09254>.
- [15] E. Reiter, S. Sripada, J. Hunter, I. Davy, Choosing words in computer-generated weather forecasts, *Artif. Intell.* 167 (2005), 137–169.
- [16] E. Reiter, S. Sripada, Should corpora texts be gold standards for NLG? in *Proceedings of the International Natural Language Generation Conference, Harriman, New York, 2002*, pp. 97–104. <https://www.aclweb.org/anthology/W02-2113/>.
- [17] R. Turner, S. Sripada, E. Reiter, I.P. Davy, (2008) Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data, in: R. Ellis, T. Allen, M. Petridis, (Eds.), *Applications and Innovations in Intelligent Systems XV*, Springer, London, 2007.
- [18] J. Kacprzyk, S. Zadrozny, Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation, *IEEE Trans. Fuzzy Syst.* 18 (2010), 461–472.
- [19] A. Ramos-Soto, A. Bugarín, S. Barro, Fuzzy sets across the natural language generation pipeline, *Prog. Artif. Intell.* 5 (2016), 261–276.
- [20] R. Turner, S. Sripada, E. Reiter, I.P.D. Davy, Using spatial reference frames to generate grounded textual summaries of georeferenced data, in *Proceedings of the 2008 International Conference on Natural Language Generation (INLG08), Association for Computational Linguistics (ACL), Salt Fork, June 12–14, 2008*. <https://www.aclweb.org/anthology/W08-1104>
- [21] R. Turner, S. Sripada, E. Reiter, Generating approximate geographic descriptions, in: E. Kraemer, M. Theune (Eds.), *Empirical Methods in Natural Language Generation, Lecture Notes in Computer Science*, vol. 5790, Springer, Berlin, Heidelberg, 2010, pp. 121–140.
- [22] S.C. Levinson, *Space in Language and Cognition: Explorations in Cognitive Diversity*, Cambridge University Press, Cambridge, vol. 5, 2003.
- [23] R. de Oliveira, S. Sripada, E. Reiter, Absolute and relative properties in geographic referring expressions, in *Proceedings of the 9th International Conference on Natural Language Generation (INLG), Edinburgh, 2016*, pp. 256–264.
- [24] P. Fisher, Sorites paradox and vague geographies, *Fuzzy Sets Syst.* 113 (2000), 7–18.

- [25] V.B. Robinson, A perspective on the fundamentals of fuzzy sets and their use in geographic information systems, *Trans. GIS.* 7 (2003), 3–30.
- [26] P. Fisher, A. Comber, R. Wadsworth, Approaches to uncertainty in spatial data, in: R. Devillers, R. Jeansoulin (Eds.), *Fundamentals of Spatial Data Quality*, Wiley, New Jersey, 2006, pp. 43–59.
- [27] P.F. Fisher, T.M. Orf, An investigation of the meaning of near and close on a university campus, *Comput. Environ. Urban Syst.* 15 (1991), 23–35.
- [28] Meteogalicia, Meteogalicia's web site, 2018. <http://www.meteogalicia.es>.
- [29] A. Ramos-Soto, N. Tintarev, R. de Oliveira, E. Reiter, K. van Deemter, Natural language generation and fuzzy sets: an exploratory study on geographical referring expression generation, in *IEEE World Congress on Computational Intelligence, 2016 IEEE International Conference on Fuzzy Systems*, Vancouver, 2016, pp. 587–594.
- [30] A. Ramos-Soto, J.M. Alonso, E. Reiter, K. van Deemter, A. Gatt, An empirical approach for modeling fuzzy geographical descriptors, in *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, Naples, 2017, pp. 1–6.
- [31] M. Delgado, M.D. Ruiz, D. Sánchez, M.A. Vila, Fuzzy quantification: a state of the art, *Fuzzy Sets Syst.* 242 (2014), 1–30.
- [32] M. Delgado, D. Sánchez, M.A. Vila, Fuzzy cardinality based evaluation of quantified sentences, *Int. J. Approx. Reason.* 23 (2000), 23–66.
- [33] J. Chamorro-Martínez, D. Sánchez, J.M. Soto-Hidalgo, P.M. Martínez-Jiménez, A discussion on fuzzy cardinality and quantification. Some applications in image processing, *Fuzzy Sets Syst.* 257 (2014), 85–101.
- [34] R. Castillo-Ortega, N. Marín, D. Sánchez, A.G. Tettamanzi, Linguistic summarization of time series data using genetic algorithms, in *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-11)*, Aix-les-Bains, France, 2011, vol. 1, pp. 416–423.
- [35] R. Castillo-Ortega, N. Marín, D. Sánchez, Linguistic query answering on data cubes with time dimension, *Int. J. Intell. Syst.* 26 (2011), 1002–1021.
- [36] F. Diaz-Hermida, M. Pereira-Fariña, J.C. Vidal, A. Ramos-Soto, Characterizing quantifier fuzzification mechanisms: a behavioral guide for applications, *Fuzzy Sets Syst.* 345 (2018), 1–23.
- [37] F. Diaz-Hermida, D.E. Losada, A. Bugarín, S. Barro, A probabilistic quantifier fuzzification mechanism: the model and its evaluation for information retrieval, *IEEE Trans. Fuzzy Syst.* 13 (2005), 688–700.
- [38] A. Ramos-Soto, E. Reiter, K. van Deemter, J.M. Alonso, A. Gatt, Meteorologists and Students: a resource for language grounding of geographical descriptors, in *Proceedings of the 11th International Conference on Natural Language Generation*, Tilburg, The Netherlands, 2018, pp. 421–425.
- [39] A. Ramos-Soto, E. Reiter, K. van Deemter, J.M. Alonso, A. Gatt, geodescriptors, 2018. <https://gitlab.citius.usc.es/alejandro.ramos/geodescriptors>.