Toward automatic generation of linguistic advice for saving energy at home

Patricia Conde-Clemente[1] Jose M. Alonso[2] · Gracian Trivino[3]

Abstract The increased demand of systems able to generate reports in natural language from numerical data involves the search for new solutions. This paper presents an adaptation of standard natural language generation methodologies to generate customized linguistic descriptions of data. Namely, we merge one of the most well-known architectures in the natural language generation research field together with our previous architecture for generating linguistic descriptions of complex phenomena. The latter is supported by the computational theory of perceptions which comes from the fuzzy sets and systems research field. We include a practical case of use dealing with the problem of inefficient consumption of energy at households. It generates natural language recommendations adapted to each household to promote a more responsible consumption. The proposal reveals opportunities of collaboration between the different research communities that are involved.

✉ Patricia Conde-Clemente
condepatricia@uniovi.es

Jose M. Alonso
josemaria.alonso.moral@usc.es

Gracian Trivino
gracian.trivino@phedes.com

[1] Departamento de Informática, Universidad of Oviedo, Edif. Dept. 1, Campus de Viesques s/n., 33204 Gijón, Asturias, Spain

[2] Centro Singular de Investigacion en Tecnoloxias da Informacion (CiTIUS), Universidade de Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Galicia, Spain

[3] Phedes Lab, Gijón, Asturias, Spain

Keywords Natural computing · Data science · Linguistic description of data · Computing with perceptions · Computational intelligence · Fuzzy sets and systems

1 Introduction

Computational models able to generate text, in natural language, from data face three main challenges (Deemter 2016): (1) Conceptualization (*What to say?*), (2) Formulation (*How to say it?*), and (3) Articulation (*Saying it*). Processing techniques allow the contextualization of data in specific domains and applying algorithms in order to extract knowledge. Report generation techniques allow representing the knowledge, previously extracted, in human friendly reports. They are expected to make use of the reader's everyday language to provide better understanding for all type of end-users (Ramos-Soto et al. 2016). These linguistic reports can be seen as a complement to other ways of knowledge representation. They actually reduce the effort of interpreting tables, graphs and charts.

In the literature, we identify two main research lines in the context of text generation from non-linguistic data, i.e., numerical and symbolic data (see Kacprzyk and Zadrożny 2010; Marín and Sánchez 2016; Ramos-Soto et al. 2016; Reiter and Dale 2000). Namely, natural language generation (NLG) for the so-called data-to-text (D2T) applications and linguistic descriptions of data (LDD) supported by Soft Computing tools. Actually, LDD can be seen as a sub-field of research inside NLG/D2T. It is expected to grow quickly in the near future.

NLG deals with generating texts that are indistinguishable from those produced by humans from numerical data, graphics or even other text. It exploits the potential of standard Data Science (Dhar 2013) allowing the better understanding of what underlies in such data. On the other hand, LDD deals with generating linguistic descriptions from numerical datasets. It is supported by Soft Computing techniques ready to exploit the tolerance for imprecision and uncertainty (vagueness; van Deemter 2009) inherent to human languages. Moreover, it has proved the ability to produce tractability, robustness and low-cost solutions when facing real-world problems (Zadeh 1994).

In recent years, the demand of automatic text generation systems has increased. For example, the central goal of the NatConsumers[1] project 2015 consists of developing an advanced and integral user-centered framework (supported by NLG techniques). It is aimed at making easier the implementation of efficient energy feedback programs in the domestic area.

The NatConsumers approach relies on a thorough characterization of EU energy consumers. In addition, it focuses on designing specific customized actions. They are tailored to each given consumer pattern and based on the use of natural language. It is worthy to remark that understanding, properly and in advance, consumer needs and expectations is essential in order to become successful when defining novel mechanisms for engaging citizens in more sustainable energy consumption habits.

Designing and developing text generation systems is only feasible in the context of a complex Software Engineering project. A key component of this type of project is the system architecture. A computational architecture that fits NatConsumers specifications is not available yet. It needs a thorough review of the scientific and technical literature to find the fundamentals to support the design of the architecture required for this project.

A complete state of the art on NLG/D2T and related software architectures is available in (Ramos-Soto et al. 2016). As main conclusion, authors remarked that there is not a general consensus about how to implement NLG systems, neither NLG/D2T systems nor LDD systems. Of course, we can take profit of valuable tools previously developed, regarding with the goals of this manuscript. For instance, we can highlight the generic architecture introduced by Reiter and Dale (2000) for designing NLG systems but also our own architecture for designing Linguistic Descriptions of Complex Phenomena (LDCP; Trivino and Sugeno 2013).

LDCP allows us to model complex phenomena, interpreting input data, and generating automatic text reports customized to the user needs. Nevertheless, we did not develop yet any architecture as complete as the one needed in NatConsumers project. Our previous publications did not consider either the need of identifying communication goals or classifying types of user. Here, we have added to LDCP some components introduced by Reiter and Dale: Communicative goal, User model, Knowledge source and Discourse history.

We have found some recent publications where Soft Computing techniques are applied together with NLG techniques. For example, in Ramos-Soto et al. (2015), we found a system that generates textual short-term weather forecasts on real prediction data, where the degree of cloudiness is characterized by fuzzy sets. Also, (Gatt and Portet 2016) addressed the problem of temporal uncertainty and its expression in NLG systems. Finally, (Gatt et al. 2016) tackled with the role of graduality and fuzziness for referring expression generation in visual scenes.

This paper is an extension of a previous work presented in Trivino and Sanchez-Valdes (2015). It deals with the automatic generation of linguistic advice for saving energy. The new contributions are as follows:

- Design and development of an adaptation of standard NLG methodologies able to generate customized linguistic descriptions.
- Solving a practical use case with the focus on generating linguistic advice about the energy consumption behavior at households. This case of use is carried out by using real databases and taxonomies. They are provided by Ariosz, the Hungarian NatConsumers partner.

Please note that the main objective of this paper is to propose the architecture suitable to NatConsumers necessities. This paper does not deal with the problem of how to evaluate the generated linguistic descriptions. To evaluate properly the use case, it is necessary to monitor the consumer behavior and this work is beyond the scope of this paper.

The rest of the manuscript is organized as follows. Section 2 contains a brief state of the art about NLG/D2T and LDD research lines. Section 3 presents the proposed architecture. Section 4 presents the case of use of the architecture applied to NatConsumers project. Section 5 shows the customized linguistic advice obtained for four households. Finally, Sect. 6 summarizes main conclusions.

2 Solutions for text generation

This section first presents a brief state of the art about NLG/D2T and LDD research lines. Then, it introduces the architecture proposed by Reiter and Dale (2000) for developing text generation systems. Finally, it also introduces our own architecture (Trivino and Sugeno 2013) for producing linguistic descriptions of complex phenomena. The interested reader is kindly referred to (Ramos-Soto et al. 2016) for a more exhaustive review on these research lines.

[1] Funded by the Horizon 2020 program of the European Commission.

2.1 Natural language generation from data

NLG/D2T is dated back to 1980s (Kittredge et al. 1986). Applications of NLG/D2T include the generation of weather reports from meteorological data (Coch 1998; Goldberg et al. 1994), the generation of reports about the state of neonatal babies from intensive care data (Portet et al. 2009), air quality reports (Busemann and Horacek 1997), etc. Nowadays, text generation from data is a hot open challenge (Cambria and White 2014; Ramos-Soto et al. 2016). Thus, several companies offer as commercial product the possibility of text generation from Big Data.

2.1.1 A generic architecture for natural language generation systems

As we mentioned above, Reiter and Dale (2000) describe a generic methodology and architectural framework that has served as inspiration for many NLG systems. Figure 1 shows the main components. Here, we summarize the main ideas behind this framework, paying special attention to the goal of our research.

The requirements analysis for the Reiter and Dale NLG architecture involves: (1) determination of system inputs; and (2) description of output texts to produce along with the additional information that may be required to produce such texts. This collection of input and output data is called Corpus. After the requirements analysis task is ended, it can be seen as an agreed-upon Corpus of representative target texts which contains the range of outputs to be expected in the system. In general, output texts express information that has been obtained from input data after performing certain computation.

After the development stage, we should obtain, from the Corpus analysis, a data structure that contains all the possible linguistic messages organized as a tree of choices. See Data structure based on Corpus in Fig. 1. This Reiter and Dale NLG architecture uses the input data to choose the most suitable message among the set of available possibilities, i.e., the Customized linguistic message is an instance of these possibilities.

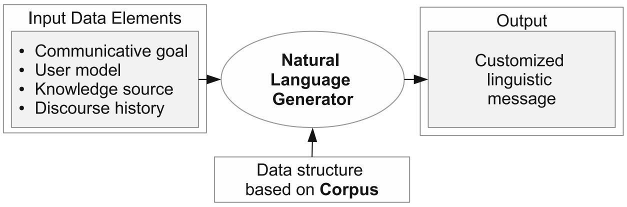


Fig. 1 Architecture for natural language generation systems proposed by Reiter and Dale (2000)

The set of input data elements are described as follows:

Communicative goal The production of linguistic messages can be viewed as a goal-driven communication process. It can be seen as an attempt to satisfy some communicative goal that the speaker has.

User model It is a characterization of the reader/hearer or intended audience for whom the text is to be generated. Among the many things that might be included in a user model is information about the user's expertise, task and preferences.

Knowledge source It may be represented in different ways depending on the application. One system may use simple tables of numbers, whereas another may use information encoded in some knowledge representation language.

Discourse history It is a model of the previous interactions between the user and the NLG system. In the simplest form, a discourse history may be just a list of the entities that have been mentioned in the discourse so far, thus providing a means of controlling the generation of anaphoric forms. More complex models would provide some characterization of the semantic content of the discourse up to the current point in order to avoid repetition.

2.1.2 The natural language generation pipeline

In order to face the complexity of the whole NLG system, Reiter and Dale divide the Text Generation Process in a pipeline with three main tasks (see Fig. 2):

Document planner It produces a document plan. The "content determination" sub-task decides what information should be communicated in the output text (*What to say?)*. In addition, the "document structuring" sub-task provides order and structure over the information to be conveyed.

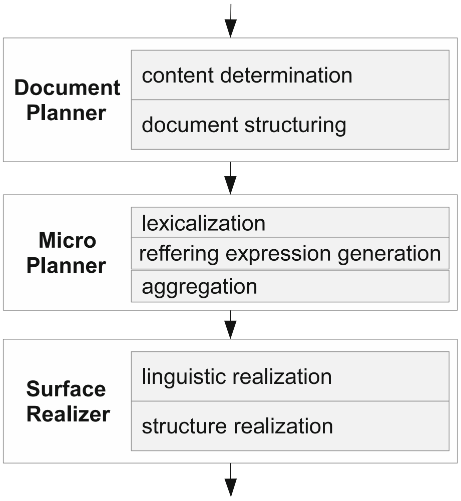


Fig. 2 The natural language generation pipeline proposed by Reiter and Dale (2000)

Micro planner It produces a text specification. The "lexicalization" sub-task solves the problem of choosing the right words (nouns, verbs, adjectives, adverbs, etc.) and syntactic structure for the generated text (*How to say it?*). Then, the "referring expression generation" sub-task produces the most suitable reference expressions to make easier the identification of entities for a reader. In addition, the "aggregation" sub-task involves the use of linguistic resources to build sentences ready to communicate several pieces of information at once. Thus, it produces a more natural text by avoiding repetitions, redundancies and so on.

Surface realizer It produces the final text. The "linguistic realization" sub-task is generally viewed as solving the problem of applying some characterization of the rules of grammar to some more abstract representation in order to produce a text (*saying it*) which is syntactically and morphologically correct. Finally, "structure realization" converts the paragraphs and sentences into the final format that will be displayed to the user.

2.2 Linguistic descriptions of data

LDD emerged with the basic concept of fuzzy linguistic summary that was established in Yager (1982) and Zadeh (1983). Typically, fuzzy linguistic summaries are based in the idea of quantified propositions which have the ability of summarizing a set of elements related to the concrete domain of a phenomenon. For example, "*Most* of the months the general consumption is *medium*" corresponds with the quantified sentence "Q of X are A", where *Most* is the quantifier Q and *medium* is the fuzzy predicate A, i.e., an attribute (feature) that characterizes the perception of the phenomenon "the general consumption" (X).

Zadeh (1999, 2002) proposed the computational theory of perceptions. This theory provides a framework to develop computational systems with the capacity of computing with the meaning of natural language expressions, i.e., with the capacity of computing with imprecise descriptions of the world in a similar way how humans do. In recent years, this concept has been extended in different ways and used for different applications, e.g., data mining (Yager 1995), database queries (Castillo-Ortega et al. 2011b; Kacprzyk et al. 2000; Kacprzyk and Zadrożny 2010), description of temporal series (Castillo-Ortega et al. 2011a; Kacprzyk et al. 2008; Kacprzyk and Yager 2001; Losada et al. 2004), comparison of time series (Castillo-Ortega et al. 2010), or the meteorology domain (Ramos-Soto et al. 2013).

2.2.1 Architecture for generating linguistic descriptions of complex phenomena

Our research for generating LDD is the so-called LDCP (Trivino and Sugeno 2013). In previous publications, we have described several phenomena using this architecture (see Fig. 3), e.g., the beauty of the double stars (Arguelles and Trivino 2013), the behavior of electricity consumers (Menendez et al. 2014), and, recently, dynamic linguistic descriptions of time series applied to self-track the physical activity (Sanchez-Valdes et al. 2016).

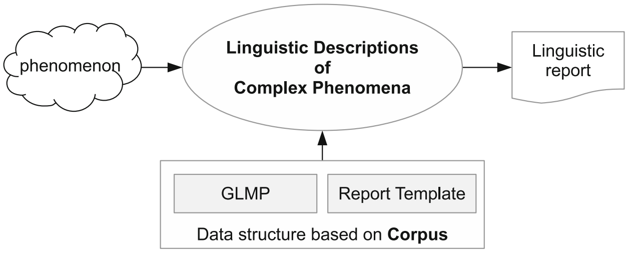

phenomenon
Linguistic Descriptions of Complex Phenomena
Linguistic report
GLMP
Report Template
Data structure based on Corpus

Fig. 3 Architecture for generating linguistic descriptions of complex phenomena

In a preliminary stage, designers collect a Corpus of natural language expressions that are typically used in the domain to describe the relevant features of the analyzed phenomenon. Then, they analyze the particular meaning of each linguistic expression in each specific situation and the user profiles to define the Granular Linguistic Model of Phenomena (GLMP) and the Report Template.

Granular linguistic model of phenomena It is a general purpose model that allows describing phenomena at different levels of granularity. GLMP is built on two main concepts, namely Computational Perception and Perception Mapping.

In general, Computational Perceptions (CP) correspond with specific parts of the analyzed phenomenon at a certain degree of granularity. As we mention above, to create a computer model of the analyzed phenomenon, the designer analyzes the everyday use of natural language about the monitored phenomenon with the aim of identifying different parts (units of information or granules) based on his/her subjective perceptions. According with Zadeh, a granule is a clump of elements which are drawn together by indistinguishability, similarity, proximity or functionality (Zadeh 1996). The GLMP handles granules by using CPs. A CP is a tuple (A, W, R) where:

A $= (a_1, a_2, \ldots, a_n)$ is a vector of linguistic expressions (words or sentences in natural language) that represents the whole linguistic domain of CP. Each component a_i is the most suitable linguistic value of CP in each situation of the phenomenon with specific granularity degree. For example, "the perception the general energy consumption" is modeled with $A = (low, medium, high)$.

W $= (w_1, w_2, \ldots, w_n)$ is a vector of validity degrees $w_i \in [0, 1]$. The validity value w_i represents the degree of each linguistic expression a_i to describe the specific input data. The sum of all validity degrees must be $\sum w_i = 1$.

R $= (r_1, r_2, \ldots, r_n)$ is a vector of relevance degrees $r_i \in [0, 1]$ assigned to each a_i in the specific context, e.g., the relevance of the linguistic expressions $A = (low, medium, high)$ is $R = (0.5, 0.5, 1)$ means the perception of *high* is considered more relevant than the other two choices. By default, all the relevance degrees take the maximum value ($r_i = 1$).

Perception Mappings (PM) are used to create and aggregate CPs. Each PM takes a set of input CPs and aggregates them into a single CP. A PM is a tuple (U, y, g, T) where:

U $= (u_1, u_2, \ldots, u_n)$ is a vector of n input CPs $u_i = (A_{ui}, W_{ui}, R_{ui})$. In the special case of first-level Perception Mappings (1PM) the inputs are values z $\in \mathbb{R}$ being provided either by sensors or obtained from a database.

y $= (A_y, W_y, R_y)$ is the output CP.

g is the aggregation function. It is divided into two functions g_W and g_R to calculate the validity W_y and relevance R_y degrees, respectively.
$W_y = g_W(W_{u_1}, W_{u_2}, \ldots, W_{u_n})$ is calculated with the validity degrees of the n input CPs. In Fuzzy Logic, many different types of aggregation functions have been developed. Indeed these aggregation functions are computational models that allow the designer to use different types of linguistic expressions. For example, in our previous publications (see Alvarez-Alvarez and Trivino 2013; Arguelles and Trivino 2013; Sanchez-Valdes et al. 2016) we implemented g_W by using a set of fuzzy rules or quantified methods (Delgado et al. 2014). In the case of 1PMs, we built g_W by using a set of membership functions.

$R_y = g_R(W_1, R_1, W_2, R_2, \ldots, W_m, R_m)$ is calculated with the validity and relevance degrees of m input CPs. This definition allows us generating more flexible reports, i.e., reports that are adapted automatically to each specific type of situation (Sanchez-Valdes et al. 2016).

T is a text generation algorithm which allows generating all the possible sentences associated with the linguistic expressions in A_y. For the sake of brevity, T is defined as a linguistic template that represents a set of possible linguistic expressions. An example of linguistic template of T is "The general energy consumption is {low | medium | high}". It allows generating three different sentences.

GLMP is a hierarchical network of CPs and PMs. The input data are introduced into the model through 1PMs which interpret these data and generate 1CPs. Then, 2PMs take several 1CPs as input and generate 2CPs, in the second level of the hierarchy. Of course, additional upper levels can be added. In such levels, 2PMs take several 2CPs as input and generate new 2CPs as output (Fig. 7, in Sect. 4.2, shows an illustrative example in relation with the use case under study).

Report template Like humans synthesize information before expressing their opinions, it is desirable that the generated reports include only the most valid and relevant information with the appropriate level of detail to each specific user. A Report Template is built by using a set of functions (programming code) and templates. The Report Generation module contains:

- The function that evaluates, sorts and selects the most interesting perceptions. Generally, it uses the validity and reliability degrees of each CP.
- The function that generates sentences. Note that it is possible to merge information from different perceptions into single sentences. Generally, this method is based on the combination of several templates.
- The function that organizes sentences in paragraphs according to the final report.

We have seen above that T is an algorithm associated with each CP that generates linguistic expressions of A. In general, the linguistic expressions generated by T are intermediate results (linguistic labels) that are dedicated to make easier the work of the designer. When the Report Generation module produces the final sentences, it could substitute these linguistic labels with more suitable linguistic expressions for each user type. E.g., expressions in another language, with different verb tense or by including emotional/sentiment content related to either warning or advice.

2.2.2 The linguistic descriptions of complex phenomena pipeline

In order to face the complexity of the whole LDCP system, we divide the Text Generation Process in a pipeline with three main tasks (Fig. 4):

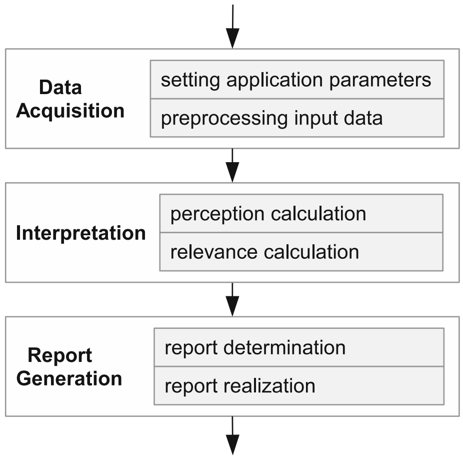


Fig. 4 The linguistic descriptions of complex phenomena pipeline

Data acquisition It takes as input both data from the specific user model and data from the phenomenon. The “setting application parameters” sub-task manages the specific user parameters. For example, it sets the user language and other details in the user model in order to select the adequate Report Template. The “preprocessing input data” sub-task manages the data from the phenomenon. It applies a set of algorithms and preprocessing techniques needed to prepare the GLMP input data, e.g., sort the data, and manage out-of-range or missing values. The result of applying these techniques must be a data structure which directly corresponds with the GLMP input.

Interpretation It takes as input the data structure created in the previous sub-task. As a result, it produces a set of linguistic expressions that are valid to describe the available data. The sub-task “perception calculation” uses the GLMP to interpret the input data, i.e., executes the network of PMs in order to calculate the validity degrees W of the CP. Then, the “relevance calculation” sub-task calculates the relevance degrees R.

Report generation It takes as input the calculated linguistic expressions. As a result, it produces the linguistic report. The “report determination” sub-task uses the Report Template to generate the linguistic report that better describes the phenomenon. Then, the “report realization” sub-task generates the output by using the Report Template.

3 Customization of linguistic descriptions

Reiter and Dale provided a generic architecture for NLG. In this section, we provide an instantiation of this generic architecture that emphasizes the representation of the meaning of the linguistic expressions. It is the result of combining both the generic architecture of NLG and the LDCP architecture.

Figure 5 shows the combination of their components: The input and output elements come from the Reiter and Dale architecture, while the Data structures (GLMP and the Report Template) come from the LDCP architecture.

Figure 6 shows the combination of their corresponding pipeline tasks. Note that the NLG pipeline tasks are focused on obtaining grammatically correct linguistic expressions, while the LDCP pipeline tasks emphasize the representation of the meaning of linguistic expressions.

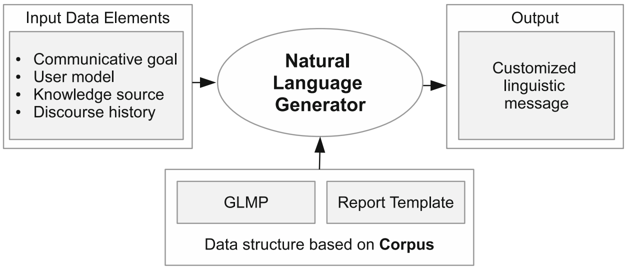


Fig. 5 Architecture for generating customized linguistic descriptions

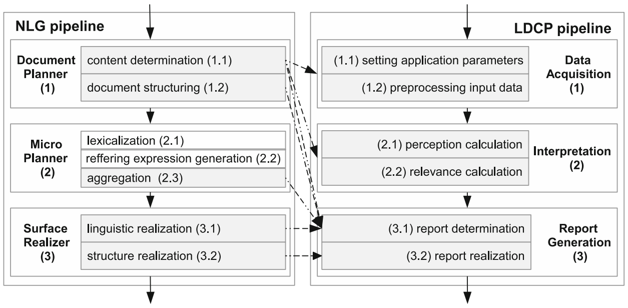

NLG pipeline
Document Planner (1)
content determination (1.1)
document structuring (1.2)
Micro Planner (2)
lexicalization (2.1)
reffering expression generation (2.2)
aggregation (2.3)
Surface Realizer (3)
linguistic realization (3.1)
structure realization (3.2)
LDCP pipeline
(1.1) setting application parameters
(1.2) preprocessing input data
Data Acquisition (1)
(2.1) perception calculation
(2.2) relevance calculation
Interpretation (2)
(3.1) report determination
(3.2) report realization
Report Generation (3)

Fig. 6 Correspondence between the NLG and LDCP pipelines

The following subsections detail: (1) how we combined the two pipelines in a new one and (2) how the resultant framework can be applied in practice from a methodological point of view.

3.1 Correspondence between NLG and LDCP pipelines

Provided the NLG generic architecture proposed by Reiter and Dale, we propose including in such architecture the components of LDCP, as follows (see Fig. 6):

Document planner In our instantiation, the LDCP (1) Data Acquisition and (2) Interpretation tasks perform the NLG (1.1) “content determination” sub-task. Also we include here the LDCP (3.1) “report determination” sub-task that evaluates, sorts and selects the most suitable perceptions. On the other hand, the LDCP (3.1) “report determination” sub-task that organizes sentences into paragraphs according to the final report corresponds with the NLG (1.2) “document structuring” sub-task.

Micro planner The LDCP (3.1) “report determination” sub-task merges information from different perceptions, thus performing the NLG (2.3) “aggregation” sub-task. Because the LDCP text generation is based on templates, this instantiation does not include the NLG (2.1) “lexicalization” and (2.2) “referring expression generation” sub-tasks.

Surface realizer The LDCP (3.1) “report determination” sub-task generates sentences, thus performing the NLG (3.1) “linguistic realization” sub-task. The LDCP (3.2) “report realization” sub-task performs the NLG (3.2) “structure realization” subtask.

3.2 Methodology

During the Requirements Analysis phase, the designer determines the input and output elements that comprise the Corpus (see Fig. 5). To accomplish this task, the designer must consider all the possibilities in the input and output elements. The different values of input data components affect the output as follows:

Communicative goal There are different versions of the Report Template in function of the Communicative goals. In addition, the designer can consider different relevance values with respect to each specific communicative goal.

User model Report Template uses specific user taxonomies as input. A more detailed characterization of users allows the designer generating more specific messages.

Knowledge source The designer must take into account both the source and the format of the data when determining the available input data to the GLMP.

Discourse history It contains the historical data and the user feedback. GLMP and Report Template use this information as input. For example, we can refer to the past and also avoid repetition in reports.

After defining the Corpus, the designer must perform the tasks in the pipeline as follows:

Data acquisition The designer must identify the source of data and the different parameters that the application can receive, e.g., goals, user models. Also, she/he must pay attention to the data structures that these parameters modify. Then, the designer must establish the mechanism required to collect the input data. In addition, she/he must define the preprocessing techniques needed to generate the GLMP input.

Interpretation The designer models the GLMP network. She/he is in charge of analyzing each expression in the Corpus of study with the aim of identifying different units of information. Then, the designer models those units of information using CPs and PMs with the appropriate degree of granularity. A key point is to identify and obtain the 1PM input data.

Report generation The designer builds the set of Report Templates. This process can be seen as a reverse engineering process whose main objective is to define several Report Templates able to generate the linguistic expressions included in the Corpus during the Requirements Analysis phase. These Report Templates should cover the different variations previously considered, e.g., goals and user models. Note that the Report Template makes use of a set of CPs from the GLMP.

In an iterative process, the designer will find out a new version of the Corpus. During iterations, the designer may find out several expressions that cannot be generated, e.g., because the lack of input data. In addition, the designer can find out new expressions that can be generated and they are relevant for the user.

During the Validation phase, the final version of this Corpus will be used to demonstrate to the user the computational system functionalities.

4 Case of use: generating linguistic advice for NatConsumers project

The architectural framework in Fig. 5 provides a general view that must be customized when we deal with implementing practical applications. In this section, we have adapted this architecture for generating linguistic advice related to energy consumption behavior at home. The next sections explain in detail the components of this application.

4.1 Input data elements

The current first stage of NatConsumers project includes undertaking surveys to consumers and experts. The main objective of this task is to establish an initial Corpus with the most appropriate advice for every situation type. Here, we present an illustrative example that is based on a very early Corpus version. The aim of this initial work is to show the possibilities of the tool to the experts in social sciences. The study and design of the final Corpus version is beyond the scope of this work.

Communicative goal It is aimed at answering the question about *what are the opportunities for saving energy?*

To do so, we consider a set of possible ways of improving the efficiency in energy consumption. For example, we can include (among others):

- To use low consumption bulbs.
- To reduce the general consumption by comparing with similar householders.
- To reduce the consumption in standby, e.g., to switch off the computer during nights.
- To move the time of the washing machine from a peak to a valley zone of consumption.
- To change the old appliances by more efficient ones.
- To improve the building thermal isolation.

User model It focuses on answering the question about *who are the consumers that need to be advised?*

It characterizes the different types of European electricity consumers, what is likely to be one of the hardest and the most challenging problems to tackle with. The goal of this characterization is generating customized advice for each kind of consumer. We have considered two taxonomies: (1) Attitudinal taxonomy based on consumers' attitudes and (2) Physical taxonomy based on physical conditions.

Attitudinal taxonomy is as follows:

Cluster 1 (*Consumption Oriented*) Averagely innovative. Not concerned about the environment. Savings are not important. Very positive about shopping.

Cluster 2 (*Modern and Passive*) Innovative. Concerned about the environment, but only at opinion level. Non-saver. Not concerned about their energy usage.

Cluster 3 (*Modern and Active*) Innovative. Strongly concerned on environmental issues, at both attitudinal and behavioral level. Cost sensitive. Concerned about their energy usage.

Cluster 4 (*Traditional Savers*) Non-innovative. Not concerned about the environment. Saving is the most important for them. They are interested in their energy consumption, but not in the environmental issues.

Physical taxonomy is as follows:

Cluster 1 Houses with maximum 2 adults with all systems, e.g., boiler, heating and air-conditioning.

Cluster 2 Houses with maximum 2 adults with some systems, e.g., boiler and air-conditioning.

Cluster 3 Houses with maximum 2 adults without any systems, e.g., without boiler, without heating and without air-conditioning.

Cluster 4 Houses with more than 2 adults or children with all systems, e.g., boiler, heating and air-conditioning.

Cluster 5 Houses with more than 2 adults or children with some systems, e.g., boiler and air-conditioning.

Cluster 6 Houses with more than 2 adults or children without any systems, e.g., without boiler, without heating and without air-conditioning.

Cluster 7 Flats with maximum 2 adults with all systems, e.g., boiler, heating and air-conditioning.

Cluster 8 Flats with maximum 2 adults with some systems, e.g., boiler and air-conditioning.

Cluster 9 Flats with maximum 2 adults without boiler, heating and air-conditioning.

Cluster 10 Flats with more than 2 adults or children with all systems, e.g., boiler, heating and air-conditioning.

Cluster 11 Flats with more than 2 adults or children with some systems, e.g., boiler and air-conditioning.

Cluster 12 Flats with more than 2 adults or children without any systems, e.g., without boiler, without heating and without air-conditioning.

Knowledge source It is related to answer the question about *what is their consumption profile?*

We consider, as source of knowledge, the data obtained thanks to the recently installed electrical energy counters in many European households. These new energy meters provide us with the following data:

Table 1 Excerpt of the dataset that contains the hourly consumption data of households

H.ID	A.ID	P.ID	Time stamp	kWh
144263	2	8	2014-01-01 00:00:00	0.155
144263	2	8	2014-01-01 01:00:00	0.13
144263	2	8	2014-01-01 02:00:00	0.125
144263	2	8	2014-01-01 03:00:00	0.145
144263	2	8	2014-01-01 04:00:00	0.115
144263	2	8	2014-01-01 05:00:00	0.1125
144263	2	8	2014-01-01 06:00:00	0.1575

Table 2 Excerpt of the dataset that contains the average hourly consumption data by Physical cluster

P.ID	Time stamp	kWh
8	2014-01-01 00:00:00	0.321
8	2014-01-01 01:00:00	0.259
8	2014-01-01 02:00:00	0.268
8	2014-01-01 03:00:00	0.283
8	2014-01-01 04:00:00	0.278
8	2014-01-01 05:00:00	0.251
8	2014-01-01 06:00:00	0.269

- Profiles of consumption of householders, i.e., the receivers of advice.
- Profile of consumption of other householders to generate comparative descriptions.
- Feedback about the obtained results. It would be highly appreciated to know if we are successfully provoking some modification of behavior.

In this first pilot for NatConsumers project, the goal is to inform consumers about their daily energy consumption and give them customized advice on how to reduce it. To achieve this goal, we evaluate the consumption of each household with respect to households with similar physical characteristics. Each household is identified by an ID and classified into one of the two taxonomies previously introduced: (1) Attitudinal and (2) Physical. The input data are structured in two datasets. Tables 1 and 2 contain samples of these datasets. The first one contains the hourly consumption data of the analyzed household. Each record includes the household ID (H.ID), the Attitudinal cluster ID (A.ID), the Physical cluster ID (P.ID), the time stamp and the hourly consumption (kWh). The second dataset contains the average consumption data of the related Physical cluster. Each record includes the Physical cluster ID (P.ID), the time stamp and the average hourly consumption (kWh).

Specifically, our system processes these inputs with the aim of generating sentences that describe: (1) the general consumption, (2) the standby consumption, (3) the consumption by parts of day and (4) the consumption with fine granularity in the different parts of the day. We define and calculate these terms as follows:

- *General consumption* It is the average consumption during certain time period. For example, Table 1 contains the hourly consumption of household with H.ID 144263 from 00 to 06 hours, 2014-01-01. The general consumption of this household is 0.134 kWh.
- *Standby consumption* It is the minimum consumption in certain days. It occurs when all activity ceases in the house. For example, the standby consumption of the household with H.ID 144263 occurs at 5 and it is 0.1125 kWh.
- *Consumption by parts of day* It is the average consumption in each part of the day. We divide the day in six parts:
  - *Dawning* between 2 and 6 A.M.
  - *Morning* between 7 and 12 A.M.
  - *Midday* between 1 and 4 P.M.
  - *Afternoon* between 5 and 7 P.M.
  - *Evening* between 8 and 9 P.M.
  - *Night* between 10 P.M. and 1 A.M.

  For example, the consumption of the household with H.ID 144263 in the Dawning is 0.131 kWh.
- Consumption with fine granularity in the different parts of the day: It is an extension of the above definition. Here, we divide each defined part in three more detailed parts. For example:
  - *Dawning first hours* between 2 and 4 A.M.
  - *Dawning central hours* between 4 and 5 A.M.
  - *Dawning last hours* between 5 and 6 A.M.

  For example, the consumption of the household with H.ID 144263 in the Dawning first hours is 0.128 kWh.

4.2 Granular linguistic model of energy consumption

Figure 7 shows the GLMP that linguistically describes the energy consumption. The GLMP input data are introduced in 1PM that aggregates the information using membership functions. The designer uses the information given in the available Corpus in order to define the modal points for each linguistic label. When linguistic labels are dynamic, the designer has to define a function capable of changing the modal points at run-time.

The reader can find below a detailed explanation of all involved PMs and CPs. Note that the input U comprises preprocessed data taken from the available datasets. The corresponding preprocess details are included in the U definition. The CPs definitions are included in the output y of its corresponding PM.

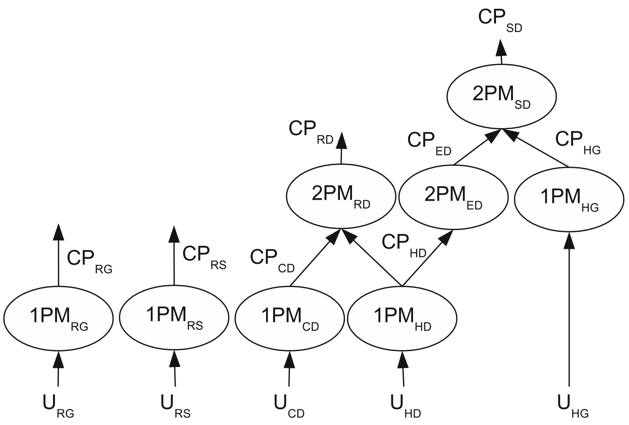


Fig. 7 GLMP that linguistically describes the energy consumption

4.2.1 General consumption

$1PM_{RG}$ compares the average general consumption of the analyzed household with respect to the general consumption in the related Physical cluster. It is defined by the tuple $(U_{RG}, y_{RG}, g_{RG}, T_{RG})$ where each component is explained as follows:

U_{RG} is the ratio between the average general consumption of the analyzed household and the average general consumption in the related Physical cluster.

y_{RG} is the $CP_{RG} = (A_{RG}, W_{RG}, R_{RG})$, where $A_{RG} =$ (*much lower*, *considerably lower*, *slightly lower*, *similar*, *slightly higher*, *considerably higher*, *almost double*, *double*, *more than double*). The relevance values are set by the designer in function of the user characteristics as follows: $R_{RG} = (1, 1, 0.8, 0.7, 0.5, 0.7, 0.8, 1, 1)$. Note that, here, we have considered that the extreme values are more relevant than the intermediate ones.

g_{RG} implements g_W by means of a set of membership functions forming a strong fuzzy partition. Notice that this kind of partition satisfies all interpretability constraints (distinguishability, coverage, etc.; Mencar and Fanelli 2008) which are required in order to build interpretable fuzzy systems, i.e., fuzzy systems easy to read and to understand by humans (Alonso et al. 2015). Thus, (see Fig. 8): {*much lower* (ml) $(-\infty, -\infty, 0.25, 0.5)$, *considerable lower* (cl) (0.25, 0.5, 0.75), *slightly lower* (sl) (0.5, 0.75, 1), *similar* (s) (0.75, 1, 1.25), *slightly higher* (sh) (1, 1.25, 1.5), *considerably higher* (ch) (1.25, 1.5, 1.75), *almost double* (ad) (1.5, 1.75, 2), *double* (d) (1.75, 2, 2.25), *more than double* (md) $(2, 2.25, \infty, \infty)$}. It is worthy to remark that triangular

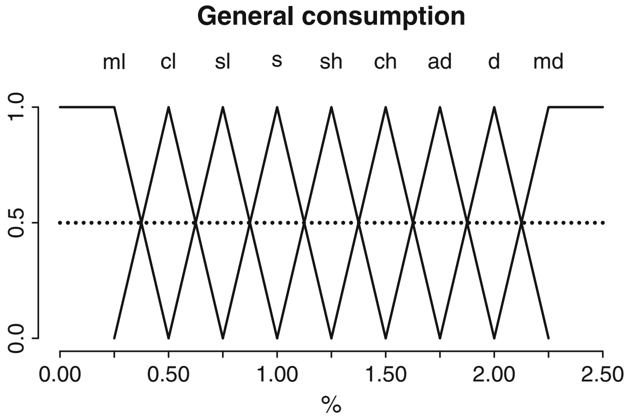


Fig. 8 Membership functions defined in g_{RG}

and trapezoidal membership functions are commonly used when building fuzzy systems. They are easy to define by experts because they satisfy interpretability constraints (normality, continuity and convexity; Mencar and Fanelli 2008). In addition, they can be efficiently and dynamically tuned in accordance with experimental data with the aim of getting a good interpretability-accuracy trade-off (Alcalá et al. 2011; Cordón et al. 2001; de Oliveira 1999). Moreover, they are easy to implement when fuzzy systems have to be embedded in low-level hardware devices.

T_{RG} is: "Your average consumption is {*much lower* | *considerably lower* | *slightly lower* | *similar* | *slightly higher* | *considerable higher* | *almost double* | *double* | *more than double*} with respect to households similar to you".

4.2.2 Standby consumption

$1\mathrm{PM}_{\mathrm{RS}}$ compares the standby consumption of the analyzed household with respect to the standby consumption in the related Physical cluster. It is defined by the tuple $(U_{\mathrm{RS}}, y_{\mathrm{RS}}, g_{\mathrm{RS}}, T_{\mathrm{RS}})$ where each component is explained as follows:

U_{RS} is the ratio of the average standby consumption of the analyzed household and the average general consumption in the related Physical cluster. The standby consumption refers to the minimum consumption that occurs when all activity ceases in the house.

y_{RS} is $\mathrm{CP}_{\mathrm{RS}} = (A_{\mathrm{RS}}, W_{\mathrm{RS}}, R_{\mathrm{RS}})$, where $A_{\mathrm{RS}} =$ (*lower, similar, higher*). $\mathrm{CP}_{\mathrm{RS}}$ takes the default value for R_{RS}.

g_{RS} implements g_W by means of a set of membership functions forming a strong fuzzy partition. The linguistic labels of A_{RS} are uniformly distributed and defined by their modal points as follows: {*lower* $(-\infty, -\infty, 0.7, 1)$, *similar* $(0.7, 1, 1.3)$, *higher* $(1, 1.3, \infty, \infty)$}.

T_{RS} is: "Your standby consumption is {*lower* | *similar* | *higher*} with respect to households similar to you".

4.2.3 Consumption in physical clusters by parts of the day

$1\mathrm{PM}_{\mathrm{CD}}$ represents the specific consumption in a given Physical cluster attending to the different parts of the day. It is defined by the tuple $(U_{\mathrm{CD}}, y_{\mathrm{CD}}, g_{\mathrm{CD}}, T_{\mathrm{CD}})$ where each component is explained as follows:

U_{CD} is a vector with the average energy consumption of the cluster by parts of the day.

y_{CD} is $\mathrm{CP}_{\mathrm{CD}} = (A_{\mathrm{CD}}, W_{\mathrm{CD}}, R_{\mathrm{CD}})$, where A is a two-dimensional array A_{CD} =((*dawning*, *morning*, *midday*, *afternoon*, *evening*, *night*),(*low*, *medium*, *high*)) that provides 6*3 different combinations of linguistic expressions.

g_{CD} implements g_W by means of a set of membership functions forming six strong fuzzy partition, one for each part of the day. These membership functions are generated dynamically using the datasets of all clusters. For example, the strong fuzzy partition of $Night$ is as follows:
low $(-\infty, -\infty, \mu - 2\sigma, \mu)$
medium $(\mu - 2\sigma, \mu, \mu + 2\sigma)$
high $(\mu, \mu + 2\sigma, \infty, \infty)$
where μ is the average and σ is the standard deviation of the energy consumption in the *Night* for all clusters. Notice that energy consumption is characterized by a Gaussian distribution.

T_{CD} is: "The energy consumption in your cluster during the {*dawning* | *morning* | *midday* | *afternoon* | *evening* | *night*} is {*low* | *medium* | *high*}".

4.2.4 Household consumption by parts of the day

$1\mathrm{PM}_{\mathrm{HD}}$ represents the specific consumption of the household attending to the different parts of the day. It is defined by the tuple $(U_{\mathrm{HD}}, y_{\mathrm{HD}}, g_{\mathrm{HD}}, T_{\mathrm{HD}})$ where each component is explained as follows:

U_{HD} is a vector with the average energy consumption of the analyzed household by parts of the day. It uses a vector calculation dual to $1\mathrm{PM}_{\mathrm{CD}}$.

y_{HD} is $\mathrm{CP}_{\mathrm{HD}} = (A_{\mathrm{HD}}, W_{\mathrm{HD}}, R_{\mathrm{HD}})$, where A is a two-dimensional array A_{HD} =((*dawning*, *morning*, *midday*, *afternoon*, *evening*, *night*), (*low*, *medium*, *high*)) that provides 6*3 different combinations of linguistic expressions. $\mathrm{CP}_{\mathrm{HD}}$ takes the default value for R_{HD}.

g_{HD} implements g_W by means of a set of membership functions forming a strong fuzzy partition for each part of the day. These strong fuzzy partitions are generated dynamically using the datasets in the related Physical cluster. We calculate these partitions in dual way to 1PM$_{\text{CD}}$.

T_{HD} is: "Your energy consumption during the {*dawning* | *morning* | *midday* | *afternoon* | *evening* | *night*} is {*low* | *medium* | *high*}".

4.2.5 Household consumption with fine granularity in the parts of the day

1PM$_{\text{HG}}$ represents the analyzed household consumption using a more fine granularity in the parts of the day. It is defined by the tuple (U_{HG}, y_{HG}, g_{HG}, T_{HG}) where each component is explained as follows:

U_{HG} is a vector with the average energy consumption of the analyzed household by parts of the day.

y_{HG} is CP$_{\text{HG}}$ = (A_{HG}, W_{HG}, R_{HG}), where A is a two-dimensional array A_{HG} =((*dawning first hours*, *dawning central hours*, *dawning last hours*, . . ., *night first hours*, *night central hours*, *night last hours*), (*low*, *medium*, *high*)) that provides 18*3 different combinations of linguistic expressions. CP$_{\text{HG}}$ takes the default value for R_{HG}.

g_{HG} implements g_W by means of a set of membership functions forming eighteen strong fuzzy partitions, one for each part of the day. These strong fuzzy partitions are generated dynamically using the datasets in the related Physical cluster. We calculate these partitions in dual way to 1PM$_{\text{CD}}$.

T_{HG} is: "The energy consumption in the related Physical cluster during the {*dawning first hours* | *dawning central hours* | *dawning last hours* | . . . | *night first hours* | *night central hours* | *night last hours*} is {*low* | *medium* | *high*}".

4.2.6 Comparing the consumption at household versus the related physical cluster by parts of the day

2PM$_{\text{RD}}$ makes comparison between the consumption of the analyzed household and its cluster attending to the different parts of the day. It is defined by the tuple (U_{RD}, y_{RD}, g_{RD}, T_{RD}) where each component is explained as follows:

U_{RD} is a vector with the household and cluster consumption by parts of the day, CP$_{\text{HD}}$ and CP$_{\text{CD}}$, respectively.

y_{RD} is CP$_{\text{RD}}$ = (A_{RD}, W_{RD}, R_{RD}), where A is a two-dimensional array A_{RD} =((*dawning*, *morning*, *midday*, *afternoon*, *evening*, *night*), (*considerably lower*, *slightly lower*, *similar*, *slightly higher*, *considerably higher*)) that provides 6*9 different combinations of linguistic expressions.

g_{RD} implements g_W by using fuzzy rules, with the usual Min–Max fuzzy reasoning mechanism. For example, one of the rules is as follows:

IF (CP$_{\text{HD}}$ household consumption in the *morning* is *high*) AND (CP$_{\text{CD}}$ cluster consumption in the *morning* is *high*) THEN (CP$_{\text{RD}}$ energy consumption in the *morning* is *similar*)

T_{RD} is: "During the {*dawning* | *morning* | *midday* | *afternoon* | *evening* | *night*} your energy consumption is {*considerably lower* | *slightly lower* | *similar* | *slightly higher* | *considerably higher*} with respect to households similar to you".

4.2.7 Energy cost by parts of the day

2PM$_{\text{ED}}$ represents the energy cost attending to the different parts of the day. It is defined by the tuple (U_{ED}, y_{ED}, g_{ED}, T_{ED}) where each component is explained as follows:

U_{ED} is CP$_{\text{HD}}$. It represents the specific consumption of the household attending to the different parts of the day.

y_{ED} is CP$_{\text{ED}}$ = (A_{ED}, W_{ED}, R_{ED}), where A is a two-dimensional array A_{ED} =((*dawning*, *morning*, *midday*, *afternoon*, *evening*, *night*), (*low*, *medium*, *high*)) that provides 6*3 different combinations of linguistic expressions.

g_{ED} implements g_W by using fuzzy rules, with the usual Min-Max fuzzy reasoning mechanism. Each fuzzy rule considers the household consumption and the energy cost to compute the household energy cost. We consider that every part of the day has the following energy cost:

– *Dawning* has a *low cost*.
– *Morning* has a *high cost*.
– *Midday* has a *high cost*.
– *Afternoon* has a *medium cost*.
– *Evening* has a *medium cost*.
– *Night* has a *high cost*.

For example, a couple of fuzzy rules are as follows:

IF (CP$_{\text{HD}}$ household consumption in the *morning* is *medium* or *high*) THEN (CP$_{\text{ED}}$ energy cost in the *morning* is *high*)
IF (CP$_{\text{HD}}$ household consumption in the *morning* is *low*) THEN (CP$_{\text{ED}}$ energy cost in the *morning* is *medium*)

T_{ED} is: “Your energy cost during the {*dawning* | *morning* | *midday* | *afternoon* | *evening* | *night*} is {*low* | *medium* | *high*}”.

4.2.8 Household shifts in the energy consumption

$\mathrm{PM}_{\mathrm{SD}}$ represents the shifts in the energy consumption attending to the energy cost and the household consumption. The goals are: (1) first to analyze the load profile; and then (2) to suggest the user how to shift its household consumption from some parts of the day to others with lower energy cost. It is defined by the tuple (U_{SD}, y_{SD}, g_{SD}, T_{SD}) where each component is explained as follows:

U_{SD} is a vector with the $\mathrm{CP}_{\mathrm{ED}}$ and $\mathrm{CP}_{\mathrm{HG}}$, the energy cost of the household by parts of the day and the household consumption by parts of the day using a fine granularity.

y_{SD} is $\mathrm{CP}_{\mathrm{SD}} = (A_{\mathrm{SD}}, W_{\mathrm{SD}}, R_{\mathrm{SD}})$, where A is a two-dimensional array A_{SD} =((*dawning*, *morning*, *midday*, *afternoon*, *evening*, *night*), (*dawning*, *morning*, *midday*, *afternoon*, *evening*, *night*)) that provides 6*6 different combinations of linguistic expressions.

g_{SD} implements g_W by using fuzzy rules with the usual Min-Max fuzzy reasoning mechanism. For example, a couple of fuzzy rules are as follows:

IF ($\mathrm{CP}_{\mathrm{ED}}$ energy cost in the *morning* is *high*) and ($\mathrm{CP}_{\mathrm{HG}}$ household consumption in the *morning first hours* is *high*) THEN ($\mathrm{CP}_{\mathrm{SD}}$ shift from the *morning* to the *dawning*).
IF ($\mathrm{CP}_{\mathrm{ED}}$ energy cost in the *morning* is *high*) and ($\mathrm{CP}_{\mathrm{HG}}$ household consumption in the *morning last hours* is *high*) THEN ($\mathrm{CP}_{\mathrm{SD}}$ shift from the *morning* to the *afternoon*)

T_{SD} generates sentences as: “You should shift part of your energy consumption from the *morning* to the *dawning*”.

4.3 Report template

Figure 9 shows the Report Template to generate linguistic advice about how to improve the daily energy consumption behavior in a specific household. The top of the report contains the data able to identify the message: the household H.ID, the analyzed period and the Physical cluster. Then, it contains a personalized greeting attending to the Attitudinal taxonomy, and the generated linguistic advice about the consumption of the household attending to three different aspects: general, specific (by parts of the day) and standby consumption.

In this example, the Report Template is as follows:

“Household ID:” {household id}
“Analyzed period: from” {initial day} “to” {final day}
“Physical cluster:” {cluster definition}
{*greeting*}
“General Consumption:” {$\mathit{salutation}_{\mathrm{RG}}$} {$T_{\mathrm{RG}}$} {$\mathit{advice}_{\mathrm{RG}}$}
“Specific Consumption:” {T_{RD}} {T_{ED}} {T_{SD}}
“Standby Consumption:” {T_{RS}} {$\mathit{advice}_{\mathrm{RS}}$}.

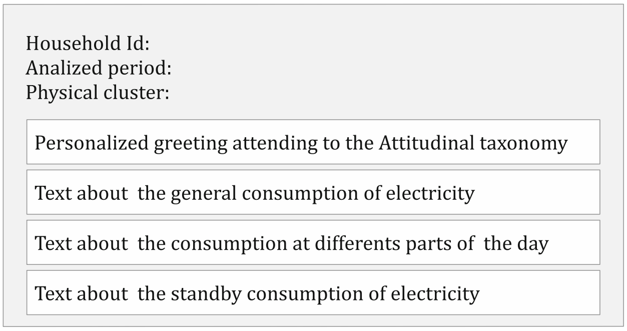

Household Id:
Analized period:
Physical cluster:
Personalized greeting attending to the Attitudinal taxonomy
Text about the general consumption of electricity
Text about the consumption at differents parts of the day
Text about the standby consumption of electricity

Fig. 9 Report template that linguistically describes the energy consumption

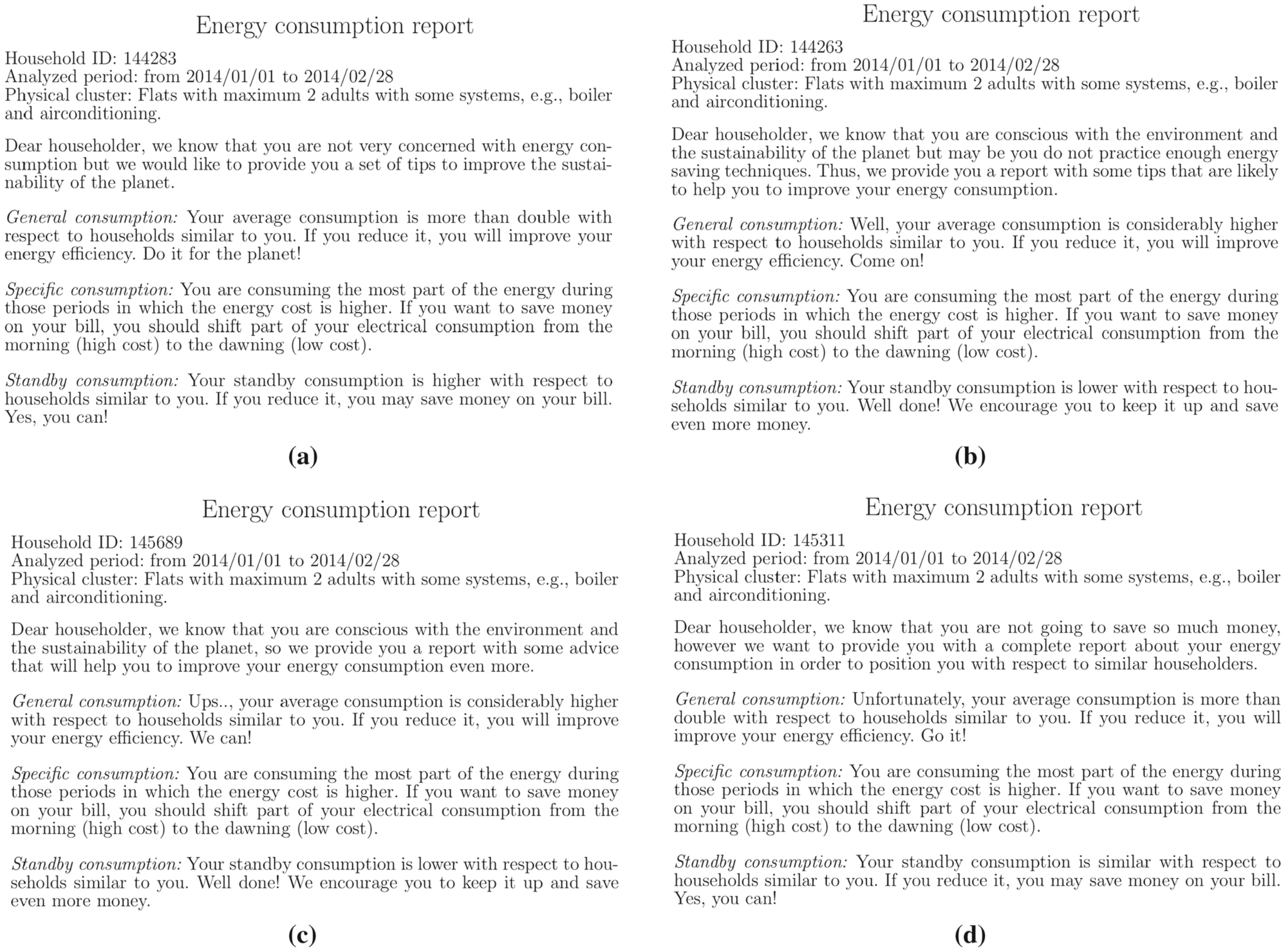
In the Report Template, the quoted texts are included directly in the final report. The components T in brackets, e.g., {T_{RG}} or {$\mathit{greeting}_{\mathrm{RG}}$}, are calls to the corresponding text generation algorithm. Here, the Report Template selects the sentence in T with the greatest validity degree and the sentence in {*greeting*} that corresponds with the associated Attitudinal cluster.

An example of the personalized greeting call is:

switch {attitudinal cluster}
case 1 “*Consumption oriented*”: {$\mathit{greeting}_1$}
case 2 “*Modern and passive*” : {$\mathit{greeting}_2$}
case 3 “*Modern and active*” : {$\mathit{greeting}_3$}
case 4 “*Traditional saver*” : {$\mathit{greeting}_4$}.

5 Discussion about illustrative examples

As we explained above, the attitudinal taxonomy offers information about the social behavior of the people living in each household. This classification allows selecting specific Report Templates for dealing with each particular type of consumer. Moreover, the Physical taxonomy offers additional information about the number of people living in the household and what type of appliances they use. This allows the designer comparing households with similar characteristics and therefore giving more realistic information about the household’s consumption.



Energy consumption report

Household ID: 144283
Analyzed period: from 2014/01/01 to 2014/02/28
Physical cluster: Flats with maximum 2 adults with some systems, e.g., boiler and airconditioning.

Dear householder, we know that you are not very concerned with energy consumption but we would like to provide you a set of tips to improve the sustainability of the planet.

General consumption: Your average consumption is more than double with respect to households similar to you. If you reduce it, you will improve your energy efficiency. Do it for the planet!

Specific consumption: You are consuming the most part of the energy during those periods in which the energy cost is higher. If you want to save money on your bill, you should shift part of your electrical consumption from the morning (high cost) to the dawning (low cost).

Standby consumption: Your standby consumption is higher with respect to households similar to you. If you reduce it, you may save money on your bill. Yes, you can!

(a)

Energy consumption report

Household ID: 144263
Analyzed period: from 2014/01/01 to 2014/02/28
Physical cluster: Flats with maximum 2 adults with some systems, e.g., boiler and airconditioning.

Dear householder, we know that you are conscious with the environment and the sustainability of the planet but may be you do not practice enough energy saving techniques. Thus, we provide you a report with some tips that are likely to help you to improve your energy consumption.

General consumption: Well, your average consumption is considerably higher with respect to households similar to you. If you reduce it, you will improve your energy efficiency. Come on!

Specific consumption: You are consuming the most part of the energy during those periods in which the energy cost is higher. If you want to save money on your bill, you should shift part of your electrical consumption from the morning (high cost) to the dawning (low cost).

Standby consumption: Your standby consumption is lower with respect to households similar to you. Well done! We encourage you to keep it up and save even more money.

(b)

Energy consumption report

Household ID: 145689
Analyzed period: from 2014/01/01 to 2014/02/28
Physical cluster: Flats with maximum 2 adults with some systems, e.g., boiler and airconditioning.

Dear householder, we know that you are conscious with the environment and the sustainability of the planet, so we provide you a report with some advice that will help you to improve your energy consumption even more.

General consumption: Ups.., your average consumption is considerably higher with respect to households similar to you. If you reduce it, you will improve your energy efficiency. We can!

Specific consumption: You are consuming the most part of the energy during those periods in which the energy cost is higher. If you want to save money on your bill, you should shift part of your electrical consumption from the morning (high cost) to the dawning (low cost).

Standby consumption: Your standby consumption is lower with respect to households similar to you. Well done! We encourage you to keep it up and save even more money.

(c)

Energy consumption report

Household ID: 145311
Analyzed period: from 2014/01/01 to 2014/02/28
Physical cluster: Flats with maximum 2 adults with some systems, e.g., boiler and airconditioning.

Dear householder, we know that you are not going to save so much money, however we want to provide you with a complete report about your energy consumption in order to position you with respect to similar householders.

General consumption: Unfortunately, your average consumption is more than double with respect to households similar to you. If you reduce it, you will improve your energy efficiency. Go it!

Specific consumption: You are consuming the most part of the energy during those periods in which the energy cost is higher. If you want to save money on your bill, you should shift part of your electrical consumption from the morning (high cost) to the dawning (low cost).

Standby consumption: Your standby consumption is similar with respect to households similar to you. If you reduce it, you may save money on your bill. Yes, you can!

(d)

Fig. 10 Illustrative examples of customized linguistic advice. **a** Household with H.ID 144283, **b** household with H.ID 144263, **c** household with H.ID 145689 and **d** household with H.ID 145311

Taking into account these two classifications, and in order to illustrate the richness of the reports, we have generated customized linguistic advice for four households. They belong to different Attitudinal clusters but the same Physical one (in order to make feasible a fair comparison among the given reports).

We analyze their consumption data from January to February 2014, paying attention to the consumption in three granularity levels: (1) General, (2) Specific (by parts of the day) and (3) Standby. Figure 10 shows the generated advice. Attending to the Attitudinal taxonomy, the household with id 144283 is Consumed Oriented, the household with id 144263 is Modern and Passive, the household with id 145689 is Modern and Active, and the household with id 145311 is Traditional Saver. Attending to the Physical taxonomy, all households live in a flat with maximum 2 adults with some systems, e.g., boiler and air-conditioning.

Linguistic advice is customized depending on the consumer model (Attitudinal taxonomy) and the energy consumption. For example, the greetings part of the template regards the Attitudinal taxonomy of the household as follows:

- Consumption Oriented (household with H.ID 144283): “Dear householder, we know that you are not very concerned with energy consumption but we would like to provide you a set of tips to improve the sustainability of the planet”.
- Modern and Passive (household with H.ID 144263): “Dear householder, we know that you are conscious with the environment and the sustainability of the planet but may be you do not practice enough energy saving techniques. Thus, we provide you a report with some tips that are likely to help you to improve your energy consumption”.
- Modern and Active (household with H.ID 145689): “Dear householder, we know that you are conscious with the environment and the sustainability of the planet, so we provide you a report with some advice that will help you to improve your energy consumption even more”.

- Traditional Saver (household with H.ID 145311): "Dear householder, we know that you are not going to save so much money, however we want to provide you with a complete report about your energy consumption in order to position you with respect to similar householders".

In both households, H.ID 144283 and 145311, the average consumption is more than double with respect to those households that are in the same Physical cluster. However, the sentences of the general consumption are adapted to the corresponding Attitudinal cluster.

Household with H.ID 144283 has the following sentence for general consumption: "Your average consumption is more than double with respect to households similar to you. If you reduce it, you will improve your energy efficiency. Do it for the planet!".

Household with H.ID 145311 has the following sentence for general consumption: "Unfortunately, your average consumption is more than double with respect to households similar to you. If you reduce it, you will improve your energy efficiency. Got it!".

Conversely, in the households with H.ID 144283 and 145689, the standby consumption is totally different, like the sentences that are generated.

Household with H.ID 144283 has the following sentence for standby consumption: "Your standby consumption is higher with respect to households similar to you. If you reduce it, you may save money on your bill. Yes you can!".

Household with H.ID 145689 has the following sentence for standby consumption: "Your standby consumption is lower with respect to households similar to you. Well done! We encourage you to keep it up and save even more money".

Since NatConsumers is an ongoing project what we have presented here is only the first pilot. Final reports will be enriched in the future with the collaboration of all partners involved in the project.

6 Conclusions

This paper presents a use case of an NLG/D2T system which, based on ideas from the general NLG architecture by Reiter and Dale, integrates elements from the computational theory of perceptions (namely the GLMP model) to determine the content of textual reports that provide linguistic descriptions of complex phenomena. The textual reports are produced by means of a Report Template close to an NLG template-based approach and take into account the communicative goal, the user model, the knowledge source and the discourse history.

The proposed case of use is carried out by using real databases and taxonomies. Currently, one of the NatConsumers tasks is the characterization of the energy consumers in Europe. This characterization will allow us to classify the different types of messages. Also, the definition of a consumer model will help us to analyze the most appropriate communication goals for each European country, climate, etc. The interested reader is kindly referred to the NatConsumers website 2015 in order to get additional information about current and future work.

In this manuscript, we have shown how to generate linguistic advice about the electric consumption: general, specific (by parts of the day) and standby. At the current level of the project development, we are working with a very initial Corpus of sentences that aim to serve as inspiration for the designers and as a mechanism for validation of the obtained initial results. Therefore, the case of use provides a demonstration about how to use the architecture. It only illustrates a first solution that is to be enhanced along the project, for example adding new elements to the Corpus.

Although there is still much to do in the field of NLG, this proposal reveals opportunities of collaboration between two previously isolated research communities (namely those ones dealing with NLG/D2T and LDD) whose collaboration is starting to take shape and is likely to yield fruitful results.

Acknowledgements We thank NatConsumers partners project for their help in performing this research. We thank especially our Hungarian partners Ariosz ltd. that have provided us with the database and the taxonomy of consumers that we have used in this work. NatConsumers project is funded by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 657672. This research was also partially funded by the Spanish Ministry of Science and Innovation under Grant FPI-MICINN BES-2012-057427 and the Spanish Ministry of Economy and Competitiveness under projects TIN2014-56633-C3-1-R, TIN2014-56633-C3-3-R and TIN2014-56967-R.

Compliance with ethical standards

Conflict of interest All authors (Patricia Conde-Clemente, Jose M. Alonso and Gracian Trivino) declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alcalá R, Nojima Y, Herrera F, Ishibuchi H (2011) Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions. Soft Comput 15(12):2303–2318

Alonso JM, Castiello C, Mencar C (2015) Interpretability of fuzzy systems: current research trends and prospects. In: Kacprzyk J, Pedrycz W (eds) Springer handbook of computational intelligence. Springer, Heidelberg, pp 219–237

Alvarez-Alvarez A, Trivino G (2013) Linguistic description of the human gait quality. Eng Appl Artif Intell 26(1):13–23

Arguelles L, Trivino G (2013) I-struve: automatic linguistic descriptions of visual double stars. Eng Appl Artif Intell 26(9):2083–2092

Busemann S, Horacek H (1997) Generating air quality reports from environmental data. In: Proceedings of DFKI workshop on natural language generation, pp 15–21

Cambria E, White B (2014) Jumping NLP curves: a review of natural language processing research. IEEE Comput Intell Mag 9(2):48–57

Castillo-Ortega R, Marín N, Sánchez D (2010) Time series comparison using linguistic fuzzy techniques. Proceedings of the 13th international conference on information processing and management uncertainty. Springer, Berlin, pp 330–339

Castillo-Ortega R, Marín N, Sánchez D (2011) A fuzzy approach to the linguistic summarization of time series. J Mult Valued Log Soft Comput 17(2–3):157–182

Castillo-Ortega R, Marín N, Sánchez D (2011) Linguistic query answering on data cubes with time dimension. Int J Intell Syst 26(10):1002–1021

Coch J (1998) Interactive generation and knowledge administration in MultiMeteo. In: Proceedings of the 9th international workshop on natural language generation, pp 300–303

Cordón O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases, vol. 19. Advances in fuzzy systems—applications and theory World Scientific Publishing, Singapore

Delgado M, Ruiz MD, Sánchez D, Vila MA (2014) Fuzzy quantification: a state of the art. Fuzzy Sets Syst 242:1–30

Dhar V (2013) Data science and prediction. Commun ACM 56(12):64–73

de Oliveira JV (1999) Semantic constraints for membership function optimization. IEEE Trans Syst Man Cybern A 29(1):128–138

Gatt A, Marín N, Portet F, Sánchez D (2016) The role of graduality for referring expression generation in visual scenes. Springer International Publishing, Cham, pp 191–203

Gatt A, Portet F (2016) Multilingual generation of uncertain temporal expressions from data: a study of a possibilistic formalism and its consistency with human subjective evaluations. Fuzzy Sets Syst 285:73–93

Goldberg E, Driedger N, Kittredge RI (1994) Using natural-language processing to produce weather forecasts. IEEE Expert 9(2):45–53

Kacprzyk J, Wilbik A, Zadrozny S (2008) Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets Syst 159(12):1485–1499

Kacprzyk J, Yager RR (2001) Linguistic summaries of data using fuzzy logic. Int J Gen Syst 30:133–154

Kacprzyk J, Yager R, Zadrożny S (2000) A fuzzy logic based approach to linguistic summaries of databases. Int J Appl Math Comput Sci 10:813–834

Kacprzyk J, Zadrożny S (2010) Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries and natural language generation. IEEE Trans Fuzzy Syst 18(3):461–472

Kittredge R, Polguére A, Goldberg E (1986) Synthesizing weather forecasts from formated data. In: Proceedings of the 11th conference on computational linguistics, pp 563–565

Losada DE, Díaz-Hermida F, Bugarín A, Barro S (2004) Experiments on using fuzzy quantified sentences in adhoc retrieval. In: Proceedings of the ACM symposium applied computing, pp 1059–1064

Marín N, Sánchez D (2016) On generating linguistic descriptions of time series. Fuzzy Sets Syst 285:6–30

Mencar C, Fanelli A (2008) Interpretability constraints for fuzzy information granulation. Inf Sci 178(24):4585–4618

Menendez C, Eciolaza L, Trivino G (2014) Generating advices with emotional content for promoting efficient consumption of energy. Int J Uncertain Fuzziness Knowl Based Syst 22(5):677–697

NatConsumers European project: (2015). http://www.natconsumers.eu. Accessed 19 Oct 2016

Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, Sykes C (2009) Automatic generation of textual summaries from neonatal intensive care data. Artif Intell 173(7–8):789–816

Ramos-Soto A, Bugarín A, Barro S (2016) On the role of linguistic descriptions of data in the building of natural language generation systems. Fuzzy Sets Syst 285:31–51

Ramos-Soto A, Bugarín A, Barro S, Díaz-Hermida F (2013) Automatic linguistic descriptions of meteorological data. A soft computing approach for converting open data to open information. In: Proceedings of the 8th Iberian conference on information systems and technologies (CISTI)

Ramos-Soto A, Bugarín AJ, Barro S, Taboada J (2015) Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. IEEE Trans Fuzzy Syst 23(1):44–57

Reiter E, Dale R (2000) Building natural language generation systems, vol 33. Cambridge University Press, Cambridge

Sanchez-Valdes D, Alvarez-Alvarez A, Trivino G (2016) Dynamic linguistic descriptions of time series applied to self-track the physical activity. Fuzzy Sets Syst 285:162–181

Trivino G, Sanchez-Valdes D (2015) Generation of linguistic advices for saving energy: architecture. In: Proceedings of the 4th international conference theory and practice of natural computing, pp 83–94

Trivino G, Sugeno M (2013) Towards linguistic descriptions of phenomena. Int J Approx Reason 54(1):22–34

van Deemter K (2009) Utility and language generation: the case of vagueness. J Philos Log 38(6):607–632

van Deemter K (2016) Computational models of referring: a study in cognitive science. MIT Press, Cambridge

Yager RR (1982) A new approach to the summarization of data. Inf Sci 28:69–86

Yager RR (1995) Fuzzy summaries in database mining. In: Proceedings of the 11th conference on artificial intelligent for applications, pp 265–269

Zadeh LA (1983) A computational approach to fuzzy quantifiers in natural languages. Comput Math Appl 9:149–184

Zadeh LA (1996) Fuzzy sets and information granularity. In: Klir GJ, Yuan B (eds) Fuzzy sets, fuzzy logic, and fuzzy systems. World Scientific Publishing Co., Inc, River Edge, NJ, pp 433–448

Zadeh LA (1999) From computing with numbers to computing with words–from manipulation of measurements to manipulation of perceptions. IEEE Trans Circuits Syst I 45(1):105–119

Zadeh LA (2002) Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. J Stat Plann Inference 105:233–264

Zadeh LA (1994) Soft computing and fuzzy logic. IEEE Softw 11(6):48–56