




Article

# A Dialogue-Act Taxonomy for a Virtual Coach Designed to Improve the Life of Elderly

César Montenegro <sup>1,†</sup>, Asier López Zorrilla <sup>2,†</sup>, Javier Mikel Olaso <sup>2,†</sup>, Roberto Santana <sup>1,\*</sup>, Raquel Justo <sup>2</sup>, Jose A. Lozano <sup>1,3</sup> and María Inés Torres <sup>2,\*</sup> 

<sup>1</sup> ISG-UPV/EHU, Faculty of Computer Science, Paseo Manuel sde Lardizabal 1, 20018 Donostia-San Sebastian, Gipuzkoa, Spain

<sup>2</sup> SPIN-UPV/EHU, Department of Electrical and Electronics, Faculty of Science and Technology, Campus de Leioa, 48940 Leioa, Bizkaia, Spain

<sup>3</sup> BCAM, Alameda de Mazarredo 14, 48009 Bilbao, Spain

\* Correspondence: roberto.santana@ehu.eus (R.S.); manes.torres@ehu.eus (M.I.T.)

† These authors contributed equally to this work.

Received: 11 June 2019; Accepted: 4 July 2019; Published: 11 July 2019

**Abstract:** This paper presents a dialogue act taxonomy designed for the development of a conversational agent for elderly. The main goal of this conversational agent is to improve life quality of the user by means of coaching sessions in different topics. In contrast to other approaches such as task-oriented dialogue systems and chit-chat implementations, the agent should display a pro-active attitude, driving the conversation to reach a number of diverse coaching goals. Therefore, the main characteristic of the introduced dialogue act taxonomy is its capacity for supporting a communication based on the GROW model for coaching. In addition, the taxonomy has a hierarchical structure between the tags and it is multimodal. We use the taxonomy to annotate a Spanish dialogue corpus collected from a group of elder people. We also present a preliminary examination of the annotated corpus and discuss on the multiple possibilities it presents for further research.

**Keywords:** dialogue systems; semantic annotation; NLU; dialogue acts

---

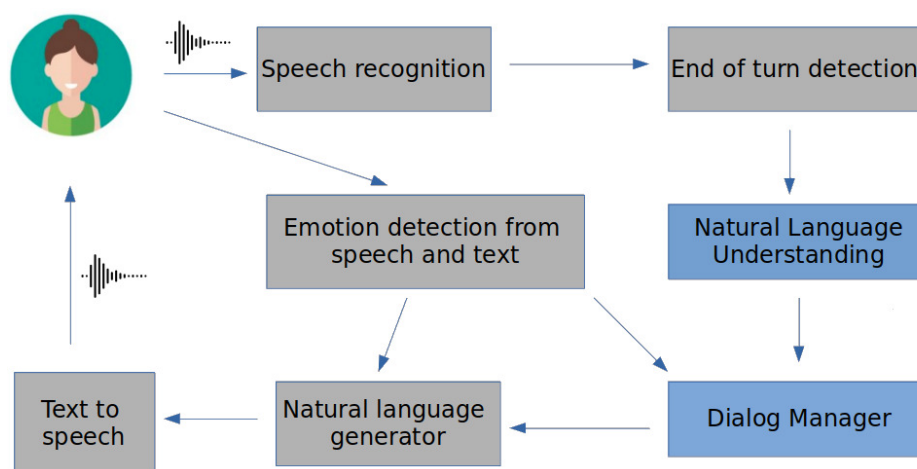
## 1. Introduction

With the advent of a new class of intelligent virtual assistants aiming to assist humans in a wide variety of tasks, new research challenges have emerged. One of these challenges is how to incorporate, to a greater extent, the emotional dimension to the interaction between the virtual assistant and the user. Another challenge is how to provide the virtual assistant with the ability to guide the user towards the achievement of its mid-term and long-term goals, which results in a new way to deal with the dialogue management. Both challenges require adapting the strategies for information exchange between the agent and the user, most notably, to adapt the way in which dialogue is conducted. Addressing these challenges requires an interdisciplinary approach involving different areas, from cognitive science to artificial intelligence.

In this paper, we focus on an important building block for the conception of a dialogue manager of an intelligent virtual assistant: the definition of a dialogue-act taxonomy for implementing the communication between the intelligent agent and the user. One particularity of our virtual agent is that the dialogue management implements a coaching model that is aimed at assisting elder people to keep a healthy and independent life. Therefore, the dialogue-act taxonomy should take into account the coaching goals of the agent as well as the particular characteristics of this population segment. Furthermore, in contrast to other applications where the conversation is guided by the user's intents, here it is the agent the one that should guide the conversation to achieve a number of coaching objectives.

A dialogue-act taxonomy for this type of systems should provide a robust platform to capture the semantics of the wide range of aspects involved in the agent-user interaction: such as providing assistance in daily tasks, suggesting health improving routines, and promoting social interactions. All these objectives require to carefully select the dialogue acts and the way they are organized.

User utterances are classified according to a previously defined set of dialogue act types, which may consist in a set of semantic units representing the translation of words into concepts. This is the work of the Natural Language Understanding (NLU) module shown in Figure 1. NLU is used to denote the task of understanding natural language coming from speech, conversation or other sources. In this work it is included in a spoken dialogue system, so it denotes the task of understanding the natural language of a human in a conversational human-machine interaction. Therefore, the definition of the act tag set or dialogue act taxonomy that will serve to label the dialogue corpus, for both the output of the NLU module and the input of the dialogue manager, is a critical step.



**Figure 1.** EMPATHIC architecture modules related to the dialogue act taxonomy definition.

Several works have addressed the question of defining dialogue act taxonomies [1–4] that will be discussed in Section 2.3. Among them, the DIT++ taxonomy [5] and the more recent ISO 24617-2 standard [6,7], which is intended to be a development of the previous one, can be considered the general methodological framework of the taxonomy defined in this paper. In this framework, our aim is to develop a taxonomy conceived for a particular application: virtual coaching designed to keep a healthy and independent life as we age [8,9]. This virtual coach is the main goal of the EMPATHIC project (Founded by the European Commission H2020-SC1-2017-RIA, grant number 769872). Studies suggest that attention to the lifestyle of the elderly can help them to maintain independent life [10]. In this application, the conversational agent is expected to guide the subject to pursue short and medium terms goals to promote healthy life style and social interaction, as well as to assist the user in the execution of daily tasks. As such, the conversational agent should be able to behave properly beyond task-specific domains.

We address the conception of such an agent from the perspective of coaching [11] which is a method that consists in accompanying, instructing, or training a person with the goal of achieving goals, or developing specific abilities. In our framework, coaching is focused on a reduced number of domains, i.e., nutrition, physical activity and social engagement. These are domains whose role is critical for keeping a healthy and independent life of elderly. As a consequence, the NLU system has to understand the user in terms of the coaching goals, thus, its output needs to be extremely related to these goals. Moreover, the dialogue-act taxonomy has to allow the dialogue manager to implement a strategy according also to the specific goals of the coaching model, in contrast with classical systems that need the decoding of the user intents. Additionally, the conversational agent is also expected to participate of more general conversation, i.e., chit-chat talk.

Thus, in brief, the main contributions of the work include the definition of a dialogue-act taxonomy aimed to represent the user utterances in the particular human-machine communication framework of the EMPATHIC project, which develops a coaching model aimed at keeping a healthy and independent life of elderly. Thus, the taxonomy allows the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. In addition, a set of real human-machine interactions in Spanish between elderly and a simulated virtual coach, i.e., through Wizard of Oz (WoZ) experiments so that the Wizard plays the role of a coach, has been annotated and discussed, providing a preliminary validation of the proposed taxonomy as well as important cues for future work in the field. All in all resulting in an original contribution in terms of language (Spanish), framework (coaching) and target population (Elderly).

The paper is organized as follows: in the next section we present the general framework of the EMPATHIC project for which the dialogue-act taxonomy has been conceived. In this section, we also briefly review the dialogue management strategy, which implements the GROW model, and discuss related work. Section 3 introduces the dialogue-act taxonomy. Section 4 describes the annotation tools and annotation procedure used to apply the introduced taxonomy to a corpus of dialogues in Spanish. It also presents an analysis of the annotated corpus for which a number of statistical measures are extracted and discussed. Finally, in Section 5 we present the conclusions of the paper and identify a number of areas for future work.

## 2. A Framework for Empathetic Conversations

### 2.1. Dialogue Acts for an Empathetic Agent

The conversational agent, as part of the project described in [8], faces several novel challenges. It will work on real time, the utterances will be automatically extracted from speech using an Automatic Speech Recognition (ASR) module, and the agent is expected to understand and produce three languages (namely Spanish, Norwegian and French, but also English, German and Italian for research support).

But notice that the results presented in this paper are for experiments conducted in Spanish. In addition, the agent will perform a variety of tasks to analyze the development of the conversations with the user.

Figure 1 illustrates just some of the main modules in the software, and the flow of information as designed. The modules *Natural Language Understanding* and *Dialogue Manager*, highlighted in blue, are the modules affected by the dialogue act taxonomy definition presented in this paper. The dialogue act set has been conceived taking into consideration these challenges and its design and description is the focus of this paper.

### 2.2. GROW Model Implemented through the Dialogue Manager

The Dialogue Management (DM) is a fundamental component of any Spoken Dialogue System (SDS). It maintains the state and manages the flow of the conversation by determining the action that the system has to perform at each agent turn. For the EMPATHIC project [9,12], we used an agenda-based management structure based on the RavenClaw [13] dialogue management framework that separates the domain-dependent and the domain-independent components of the conversation, unlike previous plan-based dialogue managers. The domain-specific aspects are defined by a dialogue task specification defined by a tree of dialogue agents. Then a domain-independent dialogue engine executes any specified task using a stack structure to control the dialog while providing reusable conversational skills, such as error recovering. This approach is suitable for dealing with complex domains while allowing the use of a relatively unconstrained natural language.

The DM and the involved strategy implement the coaching model chosen for the project. Coaching has been defined as a result-orientated, systematic process. Coaching generally uses strong questions

in order that people discover their own abilities and draw on their own resources. In other words, the role of a coach is to foster change by facilitating a coaches' movement through a self-regulatory cycle [14]. There is evidence showing that coaching interventions can be effectively applied as a change methodology [15,16]. One of the most common used coaching methodologies is the GROW Model [17]. This model provides a simple methodology and an adaptable structure for coaching sessions. Moreover, efficiency has been demonstrated in some Theoretical Behavior Change Models such as the Trans theoretical Model of Change (TTM) [18,19]. As a consequence, this coaching model has been selected for the EMPATHIC project to be integrated in the DM strategy.

A GROW coaching dialogue consists of four phases which give the name to the model: Goals or objectives, Reality, Options and Will or action plan. During the first phase, the dialogue aims at getting the specification of the objective that the user wants to achieve, for example, to reduce the amount of salt in order to diminish the related risk of hypertension. Then, this goal has to be placed within the personal context in which the user lives, and the potential obstacles need to be identified. In the next phase, the agent goal is to make the user analyse the options he/she has to achieve the objective within his/her reality. Then the final goal of the dialogue is the specification of an action plan that the user will carry out in order to advance towards goals. In this framework, the DM strategy also involves achieving the goals associated with each of the four stages. First, it will try to get a specific goal from the user, asking something like ("Would you like to improve something in your eating habits?"). Once the user provides a sentence including his/her goal, the DM will focus on the next stage. Thus, it will try to get information about the context in which the goal has to be achieved, asking something like ("How often do you usually go to the grocery?"). In this way the dialogue will be developed until all the stages are completed. This strategy, differs from classical task-oriented dialogue systems in which user asks something related to the task, and then the system tries to obtain additional information, if needed, to be able to provide as accurate a response as possible. In fact, the particular user goal and related action plan have to be agreed between the virtual agent and the user during the conversation. However, this strategy can still be correctly specified by the Ravenclaw domain dependent trees of dialogue agents mentioned above that define the domain specific aspects of the dialogue.

The EMPATHIC virtual coach is planned to deal with four coaching subdomains: nutrition [20], physical activity [21], leisure [22] and social and family engagement.

### 2.3. Related Work

While the GROW model serves as a conceptual pillar for developing the dialogue-act taxonomy, we also look to previous approaches for dialogue-act tagging.

Coding a sentence with a set of labels goes back to speech act theory of Austin [23], which has been the basis for modern data-driven dialogue act theory. Multiple different dialogue act taxonomies have been proposed to solve the task of assigning dialogue act labels to sentences. They not only differ in the precise set of tags selected, but also in characteristics such as whether the tags are exclusive, level of detail or structure.

Dialogue act taxonomies can be characterized taking into consideration different criteria, such as the following:

- Type of communication (i.e., synchronous vs. asynchronous).
- Activity type and dialogue domain.
- Type of corpora (e.g., speech dialogues, videos, chat).
- Types of speech act classification schemes.
- Dimensions (unidimensional versus multidimensional annotation).
- Annotation tools and annotation procedure.

Books, and other written forms of communication, are asynchronous methods of communication where each message is thought beforehand. This generally gives written communication a better

structure than spoken communication, where doubts, rectifications, and external factors such as noise or user speech characteristics, may result into incomplete or fuzzy messages. In this context, the PDTB [24] taxonomy was designed for annotation of discourse relations between sentences, analyzing the conjunctions used to relate them. The sense tags described in PDTB have a hierarchical structure, but they would not suit to our coaching problem since they are designed to deal with asynchronous communication.

In a human to human conversation, these problems are solved by considering the context of the conversation. Dialogue acts need to take into account whether the communication is synchronous or asynchronous. For instance, synchronous communications allow the introduction of clarification intents, where an agent may be instructed to repeat a question or formulate it in a different manner. Such type of intent tags make no sense for asynchronous methods. The dialogue-act taxonomy introduced in this paper has been conceived for synchronous communication.

While one of the most common applications of act labeling is in the context of human to human or agent to human conversation, there are other types of activities to which they have been applied [25]; for instance, they can be used for text summarization [26]. Similarly, there is a variety of platforms and domains of applications to which act labeling methods have been applied, such as social networks [26], and classification of message board posts [27]. The proposal we introduce in this paper is oriented to represent spoken communication between an agent and a human.

Another important difference between dialogue act-taxonomies are the corpora to which they are applied and from which machine learning models are commonly learned. The corpus used is, most of the time, strongly related to the domain in which the dialogue act are going to be applied, and therefore should be able to capture the particularities of the domain. Initially, available corpora were mainly created from task-oriented dialogues [28]. More recently, larger corpora have been proposed for training end-to-end dialogue systems [29,30]. In general, these large corpora are not annotated. For a survey on available corpora for dialogue systems [31] can be consulted. The corpus used in this paper has as a particular characteristic, the fact of being obtained from elderly people, a social group for which dialogues are more scarce.

Among the dialogue-act models proposed in the literature, the approach introduced in [4] presents a framework to model dialogues in conversational speech. The dialogue act taxonomy was first based on a set of tags that was used for the annotation of the discourse structure and then modified to make it more relevant for the target corpus (the Switchboard corpus [32]) and the task. However, existing dialogue act taxonomies were not designed for the scenario described in Section 2.2, where a virtual agent is the responsible for the development of the conversation. DAMSL taxonomy [33] was developed primarily for two-agent task-oriented dialogs. Nevertheless, the Empathic taxonomy has some common features with DAMSL, since the Intent dimension defined for Empathic taxonomy contains labels related to 3 out of 4 DAMSL categories (Information level, Forward Looking Function, Backward Looking Function).

Our proposal for the Empathic project has aspects in common with DIT++ [34], although they are designed for different types of interactions. On the one hand, DIT++ is based on traditional task-oriented conversations, on the other hand Empathic is based on coaching interactions, where the agent is an active member of the conversation from the point of view that the coach guides the conversation throughout the GROW model strategy. Nevertheless, many of the labels present in the taxonomy of general-purpose functions and dimension-specific functions defined in [34] can also be found in the intent label defined for Empathic. Another work relevant for our approach is the one recently published in [35], where a hierarchical schema for dialogue representation is proposed. Although the introduced scheme is specifically conceived to support computational applications, it uses a structure of linked units of intent that resembles the hierarchical structure at the core of our proposal.

The common norm for dialogue act annotation is that a single communicative function is assigned to an utterance. However, some works propose multidimensional dialogue act taxonomies in

which multiple communicative functions may be assigned to the utterances. DAMSL considers a set of exclusive group tags as different dimensions, whereas DIT++ considers a dimension in a multidimensional system, as independent, and can be addressed independently from other dimensions. In particular, a 9-dimensional annotation scheme was defined in [3]. Similarly, we use a multidimensional taxonomy which allows us to capture richer semantic information from the dialogues. Considering multiple modes of semantic information is a requirement for implementing an agent that should be able to embed the coaching objectives as part of the dialogue strategies. Without the rich information provided by multiple types of tags, it would be very difficult to guide the user to the satisfaction of the objectives, and to evaluate whether these objectives have been fulfilled. The details of this multi-modal taxonomy are described in Section 3.

Although the multidimension criteria is similar in both taxonomies, DIT++ is designed for turn labeling, and Empathic is focused in subsentence labeling. This difference forces DIT++ to separate into two dimensions different intents that can be found in the same turn as seen in the dialogue examples found in [36]. Labeling subsentences allows the Empathic taxonomy to group aspects found in the DIT++ dimensions defined in [34] (general-purpose functions, dimension-specific functions), since a turn will be split into subsentences, having only one intent label for each one, avoiding the problem of having two intent labels in the same turn.

In addition, an important effort has been carried out to define the ISO 24617-2 standard [6,7,37] that includes the 9 dimensions defined in DIT++ and reduces the number of communicative functions, which can be specific for a particular dimension or general-purpose communicative functions that can be applied in any dimension. In addition, the standard also considers different qualifiers for the certainty or the sentiment. This approach has also been included in our proposal, which can be considered, to some extent, as a reduced and GROW-driven adaptation of main characteristics of DIT++ and ISO 24617-2 for the Empathic purposes. An additional aim of the standard is to produce interoperable annotated dialogue resources. To this end, a set of dialogues from variety of corpora and dialogue annotation schemes, such as the Map task, Switchboard, TRAINs or DBOX, have been re-annotated under ISO 24617-2 scheme to build a Dialog Bank [7].

Finally, proposals for dialogue act taxonomies also differ in the annotation tools used and annotation procedures. Humans are better than machines at understanding and annotating dialogue utterances in a detailed manner, because they have more knowledge of intentional behaviour and they have richer context models [3]. So we rely on human annotation procedures to get accurate annotations instead of using automatic methods. We explain the characteristics of our annotation procedure in Section 4.

Regarding the NLU task, having a semantic representation that is both broad coverage and simple enough to be applicable to several different tasks and domains is challenging. Thus, most NLU system approaches depend on the application and the environment they have been designed for. In this way, targeted NLU systems are based on frames that capture the semantics of a user utterance or query. The semantic parsing of input utterances in NLU typically consists of three tasks: domain classification (what is the user talking about, e.g., “travel”), intent determination (what does the user want to do, e.g., book a hotel room) and slot filling (what are the parameters of this task e.g., “two bedroom suite near disneyland”) [38]. The domain detection and intent determination tasks have been typically treated as semantic utterance classification problems [39,40]. Slot filling, instead, has been treated as a sequence classification problem in which semantic class labels are assigned to contiguous sequences of words [41], which is now addressed by bidirectional LSTM/GRU models among others [42,43]. A good review of the NLU evolution is given in [44].

While the NLU employed in this work does perform intent detection and adds entity recognition, as other approaches do, the taxonomy includes intent labels specifically conceived for the GROW model, which has to fulfil additional objectives. For example, the taxonomy has to provide a relationship among the user utterances and the goals of the GROW model, which has to be agreed between user and virtual agent, and therefore be established, during the conversation, as mentioned in Section 2.2.

### 3. Characterization of the Proposed Dialogue Act Taxonomy

As discussed in previous section, the characteristics of the taxonomy must be defined according to the conversational agent needed. In our case, the agent must maintain conversations in real time about a reduced set of topics, and follow coaching strategies to guide the user. Instead of displaying a passive or merely reactive attitude, it should be pro-active and assertive, proposing different activities and topics of conversation to the elder user. In order to mitigate the difficulties in automatic labeling, due to the reasons explained, we propose a multidimensional hierarchical taxonomy to represent the relationships between the tags. Four types of labels are used, *Topic*, *Intent*, *Polarity*, and *Entity* labels.

The *Topic* label classifies the utterance in a number of classes relevant to determine the general context in which the conversation is framed. Tracking the *Topic* label will help the conversational agent to detect when the user is changing the subject of the conversation. Due to the links between the target topics the conversational agent is designed for, it is common to switch from one topic to another. Nevertheless, the DM implements a GROW-based strategy, so in this work the DM has its own goals according to the GROW model. In this framework the *Topic* label will assist the DM to associate the user utterance to both, the user and the DM goals, which have to be agreed during the interaction.

The *Intent* label classifies the utterance in classes related to the user's communicative intentions (e.g., *question*, *inform*, etc.). Our particular choice of the *Intent* labels is based on the GROW model of coaching, which is probably the best known session structure model [45]. The set of *Intent* labels we have defined is aimed to help the conversational agent to detect Goals, Realities, Obstacles and Ways forward of the particular topics the agent has been designed to deal with.

The *Polarity* label aims at representing the sentiment associated to the semantics of the user turn, which can be very relevant to provide exploitable information to dialogue managers. We distinguish between three levels of polarity: positive, neutral and negative. This label is a product of the analysis of the text, as topic and intent labels, in contrast with emotions detected from the spoken language represented in Figure 1.

The *Entity* label is different to those ones previously described in the sense that it does not serve to classify an utterance. Instead, it is applied to classify particular elements that provide specific semantic information about the Intent label. However, since the particular set of entity labels that we have selected play an important role in the semantic analysis, we consider it as a fourth modality, together with the other three. Also, entities can be useful to improve the naturalness of the conversation, when the names of the relatives are detected, or used to formulate specific proposals.

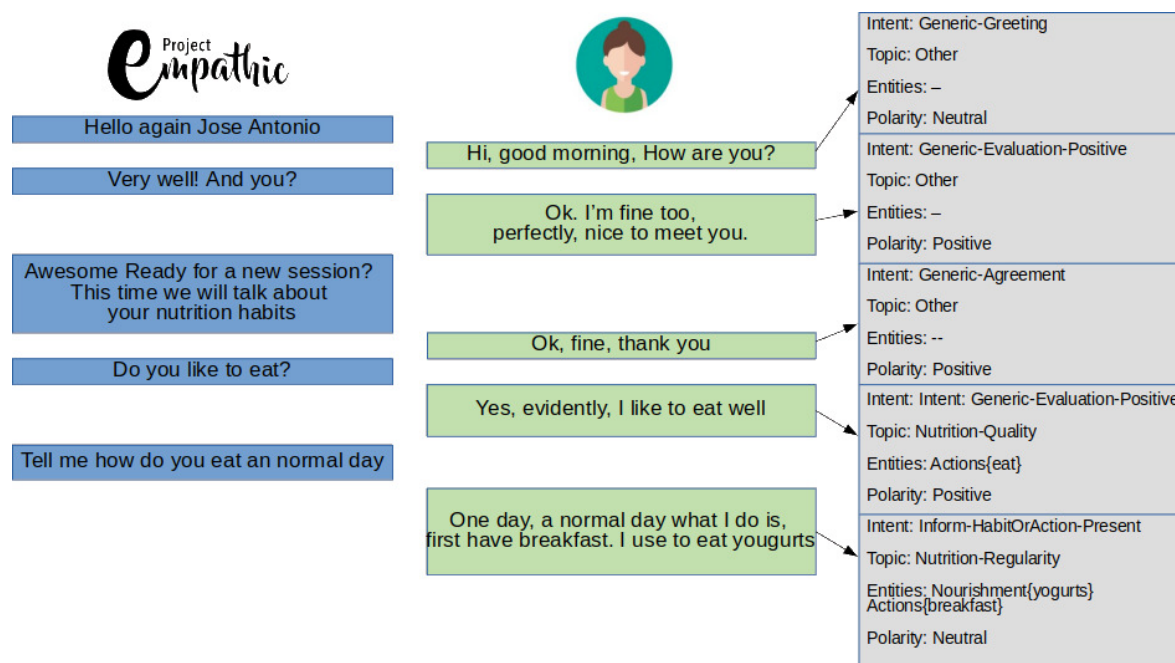
For *Topic* and *Intent* labels we propose a hierarchical structure. This means that an utterance is labeled by multiple tags that can be ordered from more general to more specific. Such labeling can be graphically represented using a tree. In this structure, the closer a label is to a leaf, the more precise it is, while the further away from the leaves, the more general. Figure A1 in Appendix A shows the topic label tag set organized as a tree. Four main groups can be recognized: *nutrition*, *sport and leisure*, *family engagement* and *other*. Each of these groups further splits into more detailed categories. Similarly, Figure A2 shows the hierarchical structure for the Intent tags. Finally, in Table 1, the entities are shown.

The rationale behind the use of hierarchical labels is to allow the agent to receive more fine-grain information when possible, but still useful less refined classification when no other choice is available. Hierarchical structures allow the experts to add more knowledge during the labeling of a dialogue corpus. In addition, when the automatic labeling model is trained, it can be less precise at the time of making predictions in those situations in which the confidence is not high enough to discriminate between two labels, selecting the parent label. This ambiguity, permits the conversational agent to guess depending on what it is expected and taking into account the other labels available. In addition, it allows the virtual coach to understand the user in terms of the system goals and topics, and thus to keep the control of the dialogue. Also, it permits the conversational agent to formulate specific questions to solve the ambiguity.

**Table 1.** List of Entity categories.

Persons	Relatives	Objects/Utensils
Actions	Nourishment	Sport and leisure
Books	Cardinal numbers	Music/Bands
Quantities	Ordinal numbers	Films/TV Series
Frequencies	Time amount	Paintings/Sculpture/Art
Diseases	Absolute dates	Places, buildings and organizations
Emotions	Relative dates	Nationalities
Meteorology		

In Figure A3, a labeling example is illustrated for the Spanish corpus. Its translation to English is shown in Figure 2. This example has been labeled by a human, and even though we have context information, sometimes it is impossible to reach the tree leaves. In the ambiguous example, it is not possible to set a more specific topic label than *nutrition*, although with context information, we could deduce that the user is referring to the little amount or variety of fruit he or she eats.



**Figure 2.** English conversation example.

#### 4. Using the Taxonomy to Get a Labelled Corpus

##### 4.1. Annotation Procedure

The proposed taxonomy was applied to label the user turn of the human-machine set of conversations acquired through the Wizard of Oz technique in the EMPATHIC project. To this end two scenarios were chosen. The first is an introductory and quite open dialogue, where the machine presents itself and asks the user about his or her hobbies. The main goal of this scenario was to make the participant feel comfortable while interacting with the simulated virtual agent. The dialogues of the second type implement a coaching session in the area of nutrition, according to the GROW model. Dialogues were acquired in three different countries with different language and culture: Spain, France and Norway. In total, 192 elder participants are interacting with the system, 72 in Spain and 60 in both France and Norway. Every user speaks with the machine in the two scenarios, and thus the final corpus will consist in 384 dialogues. Each dialogue is approximately 10 minutes long, which results in an average of 30 turns per dialogue.



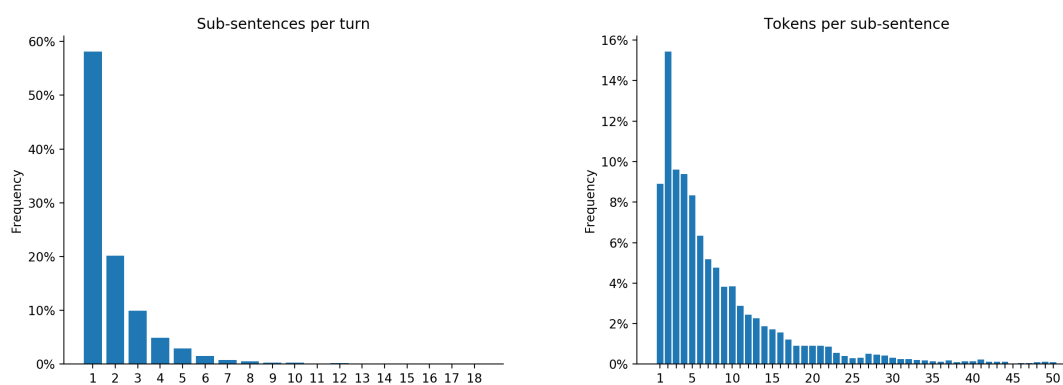
For the moment only the Spanish dialogues have been annotated according to the procedure shown in this work. To do so, 9 different annotators were instructed about the labels, the GROW model, and about the context of the project. Each of the annotators labeled roughly the same number of dialogues. Each dialogue was labeled by only one annotator. Nevertheless, all the annotators worked together to deal with doubts and disagreements, under a close supervision of the first and the second author of the paper, resulting in a collaborative annotation task. Each annotator labeled dialogues corresponding to both the introduction and the nutrition scenarios. Table 2 shows the main numbers of the annotated corpus.

**Table 2.** Description of the annotated corpus.

Characteristics	Number
Number of users	72
Number of dialogues	142
Number of turns	4522
Number of running words	72,350
Vocabulary size	5543
Number of topic labels	55
Number of intent labels	34
Number of running entities	11,113

Since more than one intent and topic can appear per turn, we asked the annotators to divide each turn into subsentences that roughly correspond to uttered clauses, so unique intent and topic labels can be assigned to each of these subsentences. To do so and to carry out the annotation procedure, we developed an annotation tool that provides a simple command-line interface. This tool shows all the user turns in a dialogue sequentially. For each of them it first asks to identify the entities. Then the annotator splits the user turn into the subsentences. Finally he or she selects, for each subsentence, the topic, intent and polarity labels. The annotators took around an hour to label each dialogue, on average.

After the annotation process, each turn was divided into 1.92 subsentences on average. The left-hand side of Figure 3 shows a histogram of the number of subsentences that resulted from the splitting of turns. On the other hand, it also shows the distribution of the number of tokens (words and punctuation marks) per subsentence. These figures show a low number of sentences per turn as well as a low number of tokens per clause or sub-sentence. These distributions are consistent to human-machine interactions where there is a significant number of user turns just consisting on two or three words that express agreement, i.e., *yes*, or disagreement, i.e., *no*.



**Figure 3.** (Left) Number of subsentences per turn. (Right) Number of tokens per subsentence.

#### 4.2. Analysis and Discussion

In this section we will show the results and statistics related to these annotations. We will first focus on general statistics of the acquired dialogues and we will then dig into the annotation results. Table 3 shows the frequency of the most frequent topics appearing in the data. The sets of frequent labels are much more reduced than the sets of all possible labels shown in Figure A1 in the Appendix A. The main reason for this is that even though the trees in Figures A1 and A2 were designed for the whole task of the EMPATHIC Project, the acquired data corresponds only to the explained two scenarios: the introduction and the nutrition scenario. As a consequence, a significant number of labels are under the *nutrition* domain whereas *hobbies* and *travelling* are associated to the welcome or introductory scenario. The label *other* includes the less frequent sub-labels as well as clauses that cannot be classified in terms of topics, such as generic agreement or disagreements. In the same way, Table 4 shows the frequency of the most frequent *Intent* labels. This table shows a significant incidence of the GROW related sub-trees. Thus the taxonomy proposed to represent the GROW model has demonstrated to be able to cover real users interactions with a Wizard who plays the role of a Virtual Coach. In the same way, the high number of generic communication tools expressing opinion, as well as agreement or disagreement, depict well spontaneous human-machine conversations. This table also shows a certain positive attitude of the participants versus the virtual agent.

**Table 3.** Frequencies of the most frequent topic labels. The sets marked with the symbol \* include all the labels under a given label, and also the cases where the annotator has not selected any sub-label.

Frequent Topic Labels	Frequency
<i>nutrition</i>	16.5%
<i>sport and leisure - hobbies</i>	5.9%
<i>sport and leisure - travelling</i>	5.8%
<i>sport and leisure - *</i>	8.1%
Other	63.7%

**Table 4.** Frequencies of the most frequent intent labels. The sets marked with the symbol \* include all the labels under a given label, and also the cases where the annotator has not selected any sub-label.

Frequent Intent Labels	Frequency
<i>generic - agreement</i>	17.4%
<i>generic - disagreement</i>	4.8%
<i>generic - evaluation/opinion</i>	18.1%
<i>generic - doubt</i>	3.3%
<i>generic - greeting</i>	4.0%
<i>generic - *</i>	6.2%
<i>GROW inform - habit or action</i>	16.7%
<i>GROW inform - objective</i>	2.5%
<i>GROW inform - obstacle</i>	2.9%
<i>GROW inform - *</i>	6.8%
<i>question</i>	3.5%
<i>other</i>	13.8%

Then, the left-hand side of Figure 4 contains the distribution of the polarity labels. As might be expected, the user is often neutral, sometimes positive and only rarely negative. This table is quite consistent with the outcomes of Table 4. Further analysis will need to be carried out to determine the correlations between these labels with the valence labels obtained through the emotion annotations based on speech and also on facial expressions.

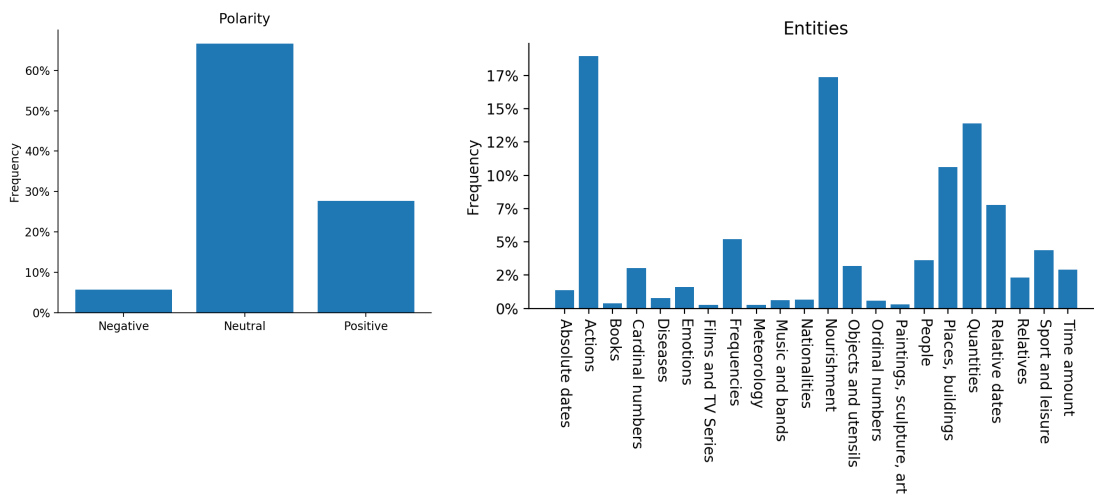


Figure 4. (Left) Distribution of the polarity in each subsentence. (Right) Distribution of the entities.

The entities were identified at the turn level. An average of 2.6 entities were labeled per user turn. The right-hand side of Figure 4 shows the frequency of each of the entities. This figure shows a significant occurrence of entities that correspond to user utterances developing the coaching model proposed by the Wizard, such as *Nourishment* and *Actions*.

In Figure 5, we illustrate the relationship between topics and intents, by means of Sankey diagram ([https://en.wikipedia.org/wiki/Sankey\\_diagram](https://en.wikipedia.org/wiki/Sankey_diagram)). In this and the following figures, the most representative labels (in terms of appearance in the labeled conversations) of two label groups face each other. The flows that connect the labels from one side with the other, represent the amount of sentences that are labeled with the two connected labels. The labels that are not representative enough, are included in the parent label appended with a star. Also, only labels down to the second level of depth are contemplated, any deeper label is included in its second depth level parent node. In order to help understanding visually the tree structure, all the labels pending from a first level node will have the same color.

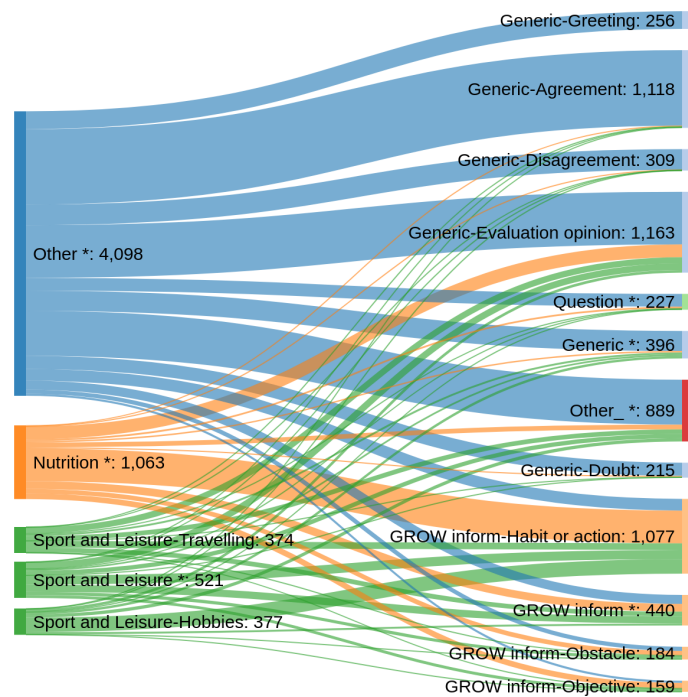


Figure 5. Relationship between intent and topic labels.

Figure 5 also allows us a better understanding of the relation between both labelling as well. The most frequent *Topic* label is *other* with a 63.86% of the sentences. But it shows that this high percentage corresponds to general conversation, greetings, or answers to questions from the Virtual coach as seen in the flows that connect the *other* topic label with all the *generic* family of labels represented in blue. As the second session was planned to be about *Nutrition*, the Virtual coach had to ask about the user’s habits. Therefore, a 16.69% of the intent labels are *GROW inform-habit or action*, what it was not expected was to have as much as *nutrition* habits explained as *sport and leisure-family habits*. Other intent labels share this particularity of having as much relation with *nutrition* as with *sport and leisure* as *generic-evaluation/opinion* and the *GROW inform* label family in light orange.

In Figures 6 and 7, Sankey diagrams are used to represent the relation between the entities and the intent and topic labels, respectively. For the sake of clarity these figures only represent the first level of *Intends* and *Topics* trees. Figure 6 shows how sentences that have *Nourishment* entities are often *GROW inform* sentences, what is consistent with the conclusions obtained from Figure 5 about the relationship between the *nutrition* *Topic* and the *GROW inform* labels. Also other entities are useful to inform about *nutrition* habits like *Actions*, *Quantities* and *Relative dates* for example.

The relations deduced in Figures 5 and 6, are reinforced with Figure 7, where we can also see the relation of *Nutrition* topic with *Nourishment*, *Quantities* and *Actions* entities among others.

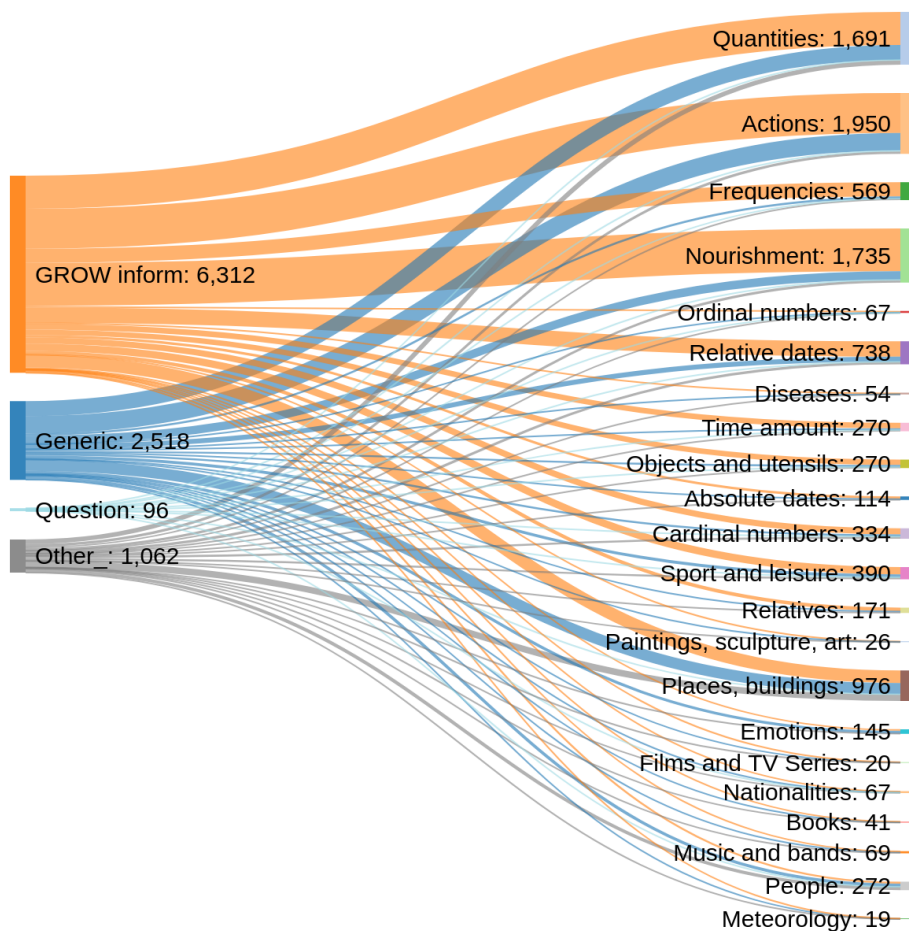
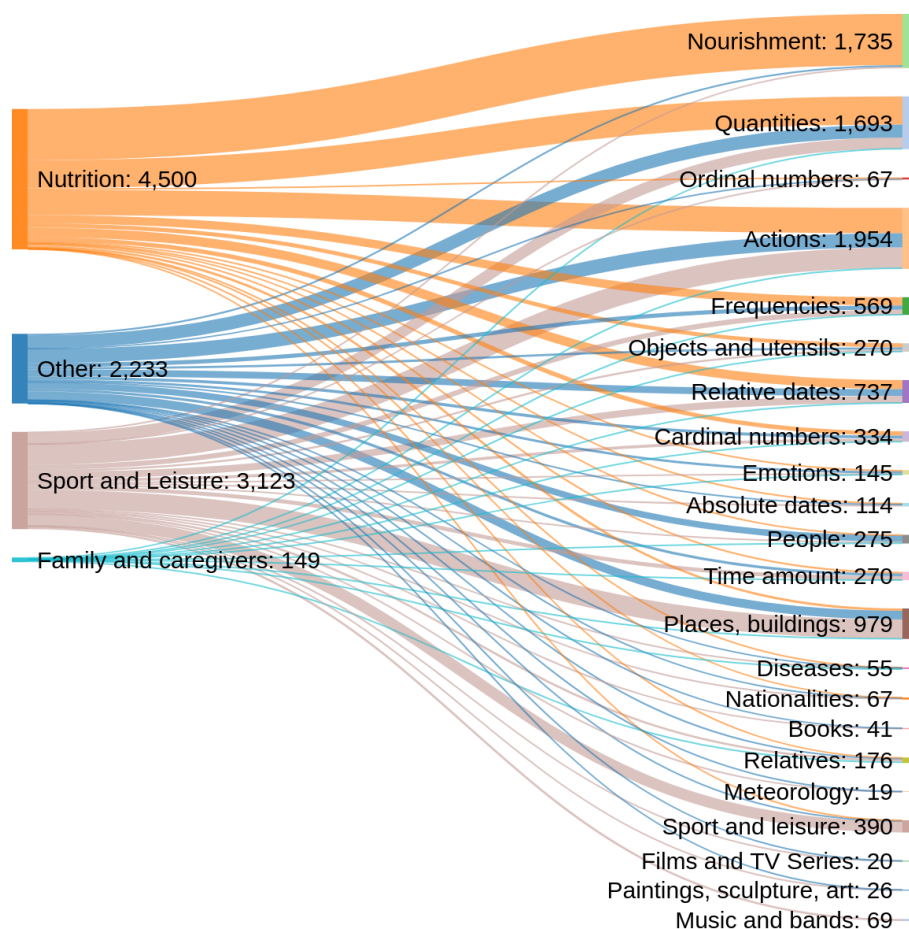


Figure 6. Relation between intent and entities.



**Figure 7.** Relation between topic and entities.

The analysis presented and discussed in this section shows how a taxonomy developed from a theoretical coaching model, such as the GROW one proposed in this work, can be followed by real users interacting with the simulated agent. The annotated data show that the taxonomy is capable of fully coverage of the spontaneous utterances of the participants in terms of concepts, topics, communicative intends that are useful to provide meaningful information to the DM according to the the goals to be developed by the system, to a certain level of granularity, but still useful. In the same way, the distribution of the selected entities seems also to agree the goals of the developed scenarios. Thus, the outcomes of the data annotation could asses the hierarchy proposed, to some extent. On the other hand, the positive agreements and opinions, as well as the polarity distributions, seem to show a relaxed and positive attitude towards the Wizard, who played well the role of a coach. All in all these data seem to support the full procedure of the WoZ recordings.

## 5. Conclusions

In this paper we have introduced a dialogue act taxonomy that has among its distinctive characteristics the capacity for supporting communication based on a coaching strategy, a hierarchical structure between the tags, and the fact of being multi-modal.

The coaching strategy is essential within the framework of EMPATHIC since it directly addresses the need to implement a pro-active agent that provides assistance and counseling to the elderly users, and drives the dialogue with the intention of reaching coaching goals. This is an important difference to other approaches such as task-oriented dialogue systems and chit-chat implementations.

The hierarchical structure allows us to capture varying degrees of semantic information from the utterances. Having different taxonomies for topics and intents allow the system a very rich semantic

representation of the dialogues that provides more flexibility for the design of dialogue managing strategies. Combined, these characteristics make our proposal significantly different to previous dialogue act taxonomies and a very relevant proposal for the implementation of virtual agents.

Another important contribution from our work is to provide one of the first analysis of an annotated corpus constructed from 142 interactions between elder people and visual agents. This corpus is precious because it covers a population usually neglected in similar studies, mainly due to the difficulties involved in accessing to elderly and face them with the required technologies. We emphasize that the usefulness of the corpus, and of the obtained annotations, goes beyond the implementation of the virtual coach. Although the results shown in this paper are limited to Dialogues in Spanish, there is ongoing work in the completion of annotated dialogues for the other two languages. Thus, further language and cultural comparisons will be achieved.

The validation of the introduced taxonomy will require the application of a classification strategy able to label the dialogues using the introduced sets of tags. The taxonomy could be indirectly evaluated in terms of the performance of the dialogue manager that uses it. On the other hand, while hierarchical multi-modal taxonomies, as the one we have introduced, are richer and provide much flexibility for the implementation of dialogue managing strategies, they are also challenging for typical machine learning methods. For instance, hierarchical multi-label classification is more difficult than traditional multi-class classification problems. Similarly, using topic label information for implementing specialized coaching scenarios and switching between them is not a trivial task. Nevertheless, we consider that the taxonomy introduced in this paper provides us with a good set of tools to face these challenges.

We foresee a number of ways in which the annotated data can be valuable for machine learning applications. While its size is relatively small (4522 turns for the Spanish corpus), this data is sufficient to refine machine learning models that had been trained using larger, more general, dialogue corpora. In addition, as part of the EMPATHIC there will be annotated data for other languages. This fact points to the feasibility of obtaining a larger corpus by translating all dialogues to a base language. Moreover, it opens the possibility of investigating transfer learning strategies in a multi-lingual framework. We have already obtained preliminary results on the application of parallel corpora for training dialogue classifiers (Montenegro et al. [46]). Finally, the annotated corpora is valuable itself due to the particular characteristics of the task as well as of the elderly population from which it has been obtained.

Finally, the analysis of the annotation data discussed in this work let to conclude that the taxonomy developed from a theoretical coaching model, such as the GROW model, has been able to provide fully coverage of the spontaneous utterances of real users interacting with a simulated agent who plays the role of a Virtual Coach. Moreover, this analysis shows significant frequencies of GROW related *Topic*, *Intends* and *Entities* labels. All in all, these outcomes could also provide a preliminary validation of the taxonomy proposed.

**Author Contributions:** R.J., R.S., J.A.L. and M.I.T. designed the project, C.M., A.L.Z. and J.M.O. developed the semantic structures, and all authors contributed equally and significantly in writing this article.

**Funding:** The research presented in this paper is conducted as part of the project EMPATHIC that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769872. The authors would also like to thank the support by the Basque Government through the project IT-1244-19.



**Conflicts of Interest:** The authors declare no conflict of interest.

Appendix A.

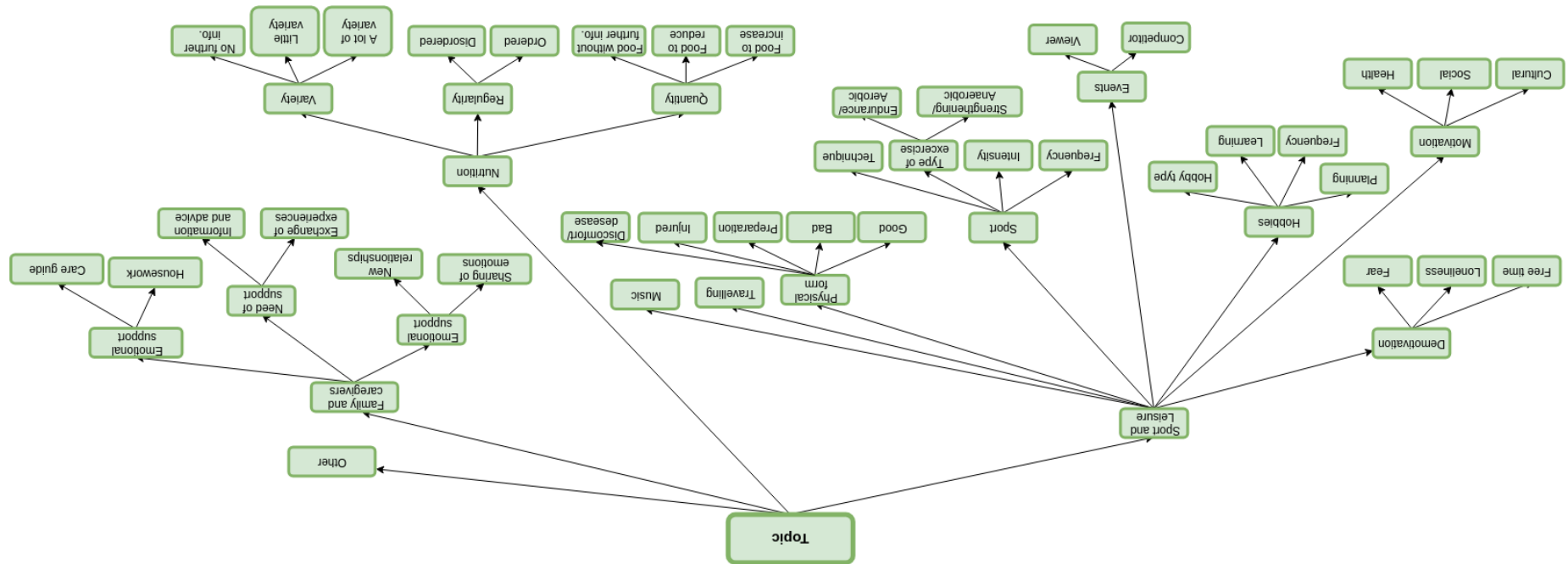


Figure A1. Topic label tree.

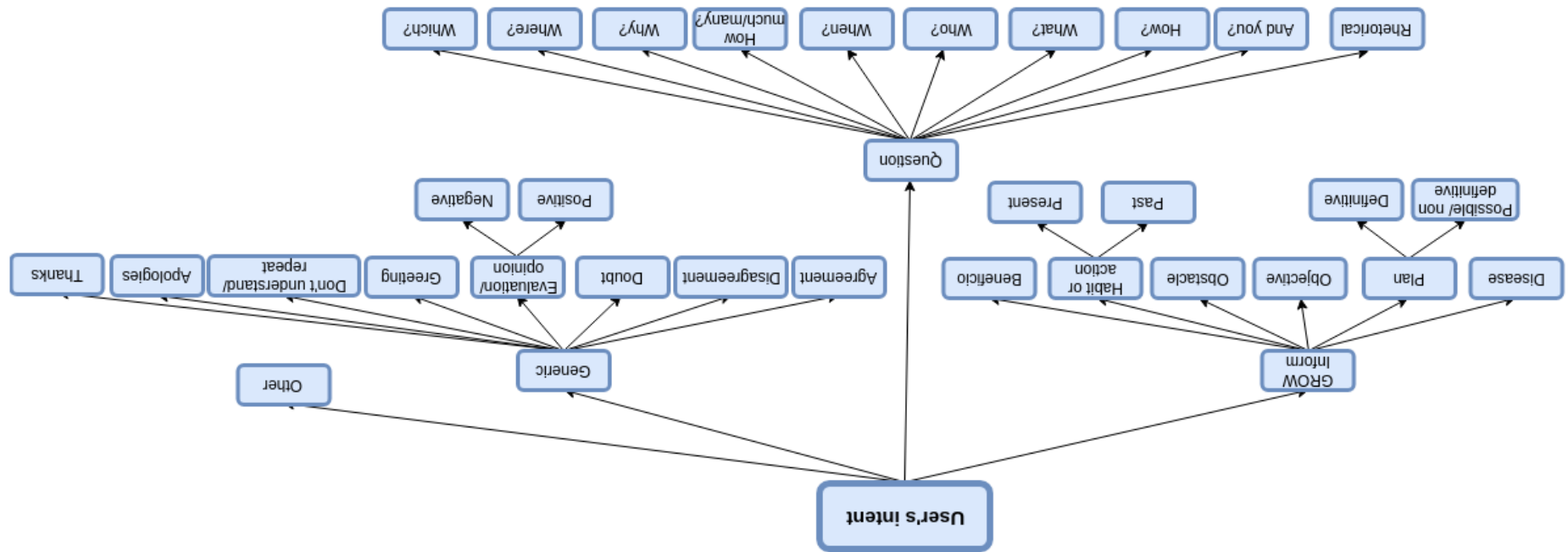


Figure A2. Intent label tree.



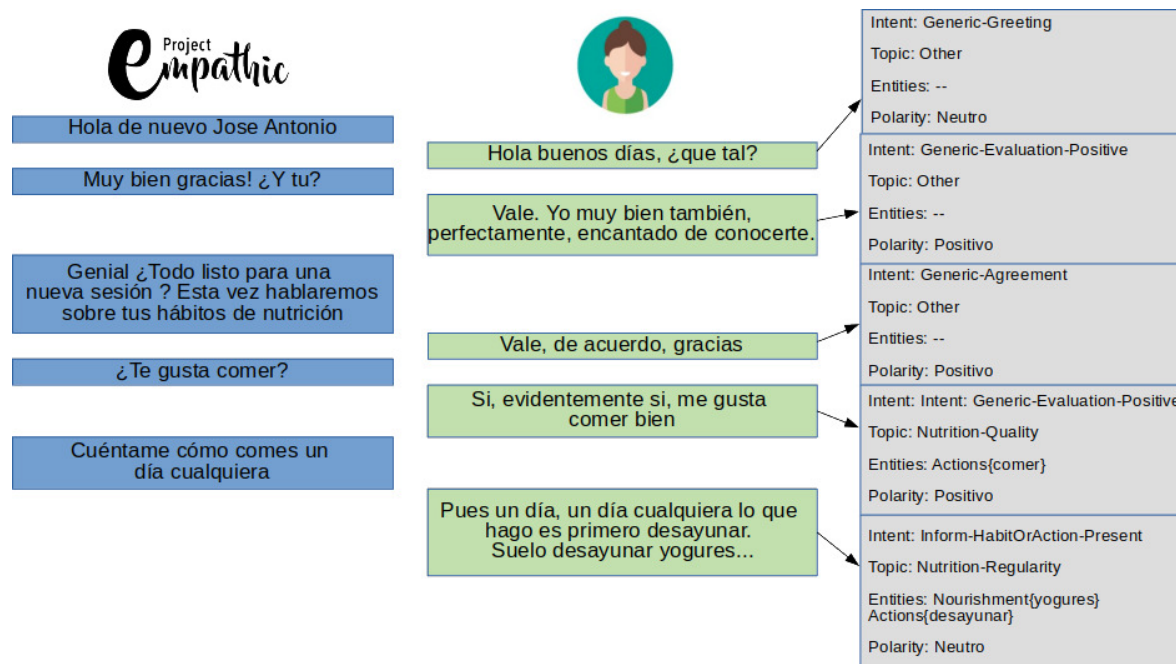


Figure A3. Spanish conversation example.

## References

- Lampert, A.; Dale, R.; Paris, C. Classifying speech acts using verbal response modes. In Proceedings of the Australasian Language Technology Workshop 2006, Sydney, Australia, 30 November–1 December 2006; pp. 34–41.
- Traum, D.R. 20 questions on dialogue act taxonomies. *J. Semant.* **2000**, *17*, 7–30. [CrossRef]
- Bunt, H.; Alexandersson, J.; Choe, J.W.; Fang, A.C.; Hasida, K.; Petukhova, V.; Popescu-Belis, A.; Traum, D.R. ISO 24617-2: A semantically-based standard for dialogue annotation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 2012; pp. 430–437.
- Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Ess-Dykema, C.V.; Meteer, M. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* **2000**, *26*, 339–373. [CrossRef]
- Bunt, H. The DIT++ taxonomy for functional dialogue markup. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, 10–15 May 2009.
- Petukhova, V.; Bunt, H. The coding and annotation of multimodal dialogue acts. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, 23–25 May 2012; pp. 430–437.
- Bunt, H.; Petukhova, V.; Malchanau, A.; Wijnhoven, K.; Fang, A. The DialogBank. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016; Chair, N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al. Eds.; European Language Resources Association (ELRA): Paris, France, 2016.
- López Zorrilla, A.; Velasco Vázquez, M.D.; Irastorza, J.; Olaso Fernández, J.M.; Justo Blanco, R.; Torres Barañano, M.I. EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly. *Proces. Del Leng. Nat.* **2018**, *61*, 167–170.
- Torres, M.L.; Olaso, J.M.; Montenegro, C.; Santana, R.; Vazquez, A.; Justo, R.; Lozano, J.A.; Schloegl, S.; Chollet, G.; Dugan, N.; et al. The EMPATHIC Project: Mid-term Achievements. In Proceedings of the 12th Conference on Pervasive Technologies Related to Assistive Environments Conference (PETRA-19), Rhodes, Greece, 5–7 June 2019.

10. Willcox, D.C.; Scapagnini, G.; Willcox, B.J. Healthy aging diets other than the Mediterranean: A focus on the Okinawan diet. *Mech. Ageing Dev.* **2014**, *136*, 148–162. [[CrossRef](#)] [[PubMed](#)]
11. Graham, A. Behavioural coaching—The GROW model. In *Excellence in Coaching: The Industry Guide*, 2nd ed.; Jonathan, P., Ed.; Kogan Page: London, UK, 2006; pp. 83–93.
12. Torres, M.I.; Olaso, J.M.; Glackin, N.; Justo, R.; Chollet, G. A Spoken Dialogue System for the EMPATHIC Virtual Coach. In Proceedings of the International Workshop on Spoken Dialog System Technology (IWSDS), Singapore, 14–16 May 2018.
13. Bohus, D.; Rudnicky, A.I. The RavenClaw dialog management framework: Architecture and systems. *Comput. Speech Lang.* **2009**, *23*, 332–361. [[CrossRef](#)]
14. Grant, A.M. The impact of life coaching on goal attainment, metacognition and mental health. *Soc. Behav. Personal.* **2003**, *31*, 253–263. [[CrossRef](#)]
15. Theeboom, T.; Beersma, B.; van Vianen, A.E. Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. *J. Posit. Psychol.* **2014**, *9*, 1–18. [[CrossRef](#)]
16. Jones, R.; Woods, S.; Guillaume, Y. The effectiveness of workplace coaching: A meta-analysis of learning and performance outcomes from coaching. *J. Occup. Organ. Psychol.* **2015**, *89*, 249–277. [[CrossRef](#)]
17. Whitmore, J. *Coaching for Performance: Growing Human Potential and Purpose: The Principles and Practice of Coaching and Leadership*; Nicholas Brealey Publishing: London, UK, 2009.
18. Passmore, J. An integrated model of goal-focused coaching: An evidence-based framework for teaching and practice. *Int. Coach. Psychol. Rev.* **2012**, *7*, 146–165.
19. Passmore, J. Motivational Interviewing—A model for coaching psychology practice. *Coach. Psychol.* **2011**, *7*, 35–39.
20. Sayas, S. *Dialogues on Nutrition*; Technical Report DP1, Empathic Project; Internal Documents: Tampere, Finland, 2018.
21. Sayas, S. *Dialogues on Physical Exercise*; Technical Report DP2, Empathic Project; Internal Documents: Tampere, Finland, 2018.
22. Sayas, S. *Dialogues on Leisure and Free Time*; Technical Report DP3, Empathic Project; Internal Documents: Tampere, Finland, 2018.
23. Austin, J.L. *How to do Things with Words*; Oxford University Press: Oxford, UK, 1975.
24. Prasad, R.; Dinesh, N.; Lee, A.; Miltasakaki, E.; Robaldo, L.; Joshi, A.K.; Webber, B.L. The Penn Discourse TreeBank 2.0. In Proceeding of the 6th Language Resources and Evaluation Conference, Marrakech, Morocco, 28–30 May 2008.
25. Popescu-Belis, A. *Dialogue Acts: One or More Dimensions*; ISSCO Work: Alexandra, New Zealand, 2005; p. 62.
26. Zhang, R.; Li, W.; Gao, D.; Ouyang, Y. Automatic twitter topic summarization with speech acts. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 649–658. [[CrossRef](#)]
27. Qadir, A.; Riloff, E. Classifying sentences as speech acts in message board posts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 748–758.
28. Anderson, A.H.; Bader, M.; Bard, E.G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; et al. The HCRC map task corpus. *Lang. Speech* **1991**, *34*, 351–366. [[CrossRef](#)]
29. Lowe, R.; Pow, N.; Serban, I.; Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv* **2015**, arXiv:1506.08909.
30. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv* **2018**, arXiv:1801.07243.
31. Serban, I.V.; Lowe, R.; Henderson, P.; Charlin, L.; Pineau, J. A survey of available corpora for building data-driven dialogue systems. *arXiv* **2015**, arXiv:1512.05742.
32. Godfrey, J.J.; Holliman, E.C.; McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of the ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, USA, 23–26 March 1992; Volume 1, pp. 517–520.
33. Allen, J.; Core, M. Draft of DAMSL: Dialog Act Markup in Several Layers. Available online: <http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/> (accessed on 11 June 2019).
34. Bunt, H. The DIT++ taxonomy for functional dialogue markup. In Proceedings of the AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts, Budapest, Hungary, 10–15 May 2009; pp. 13–24.

35. Pareti, S.; Lando, T. Dialog Intent Structure: A Hierarchical Schema of Linked Dialog Acts. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, 12 May 2018.
36. Keizer, S.; Bunt, H.; Petukhova, V. Multidimensional dialogue management. In *Interactive Multi-Modal Question-Answering*; Springer: Cham, Switzerland, 2011; pp. 57–86.
37. Bunt, H.; Petukhova, V.; Traum, D.; Alexandersson, J. Dialogue Act Annotation with the ISO 24617-2 Standard. In *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*; Dahl, D.A., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 109–135.
38. Tur, G.; DeMori, R. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*; John Wiley and Sons: Hoboken, NJ, USA, 2011.
39. Yaman, S.; Deng, L.; Yu, D.; Wang, Y.; Acero, A. An Integrative and Discriminative Technique for Spoken Utterance Classification. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1207–1214. [[CrossRef](#)]
40. Heck, L.; Hakkani-Tür, D. Exploiting the Semantic Web for unsupervised spoken language understanding. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2–5 December 2012; pp. 228–233. [[CrossRef](#)]
41. Wang, Y.; Acero, A. Discriminative models for spoken language understanding. In Proceedings of the INTERSPEECH 2006—ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006; ISCA: Woodbridge, ON, Canada, 2006.
42. Hakkani-Tür, D.; Tür, G.; Çelikyilmaz, A.; Chen, Y.; Gao, J.; Deng, L.; Wang, Y. Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. In Proceedings of the Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; Morgan, N., Ed.; ISCA: Woodbridge, ON, Canada, 2016; pp. 715–719. [[CrossRef](#)]
43. Vukotic, V.; Pintea, S.; Raymond, C.; Gravier, G.; van Gemert, J.C. One-Step Time-Dependent Future Video Frame Prediction with a Convolutional Encoder-Decoder Neural Network. *arXiv* **2017**, arXiv:1702.04125.
44. Tur, G.; Celikyilmaz, A.; He, X.; Hakkani-Tür, D.; Deng, L. Deep Learning in Conversational Language Understanding. In *Deep Learning in Natural Language Processing*; Deng, L., Liu, Y., Eds.; Springer: Singapore, 2018; pp. 23–48. [[CrossRef](#)]
45. Grant, A.M. Is it time to REGROW the GROW model? Issues related to teaching coaching session structures. *Coach. Psychol.* **2011**, *7*, 118–126.
46. Montenegro, C.; Santana, R.; Lozano, J.A. Data generation approaches for topic classification in multilingual spoken dialog systems. In Proceedings of the 12th Conference on Pervasive Technologies Related to Assistive Environments Conference (PETRA-19), Rhodes, Greece, 5–7 June 2019; ACM: New York, NY, USA, 2019.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).