Book Title
Book Editors
IOS Press, 2003

Some Foundational Linguistic Elements for QA Systems: an Application to E-government Services

Farida Aouladomar [1]
IRIT, France.

Abstract. Time saving and time flexibility of eGovernment procedures is more attractive than face-to-face services to citizens. Citizens may interact with government via emails, search administrative information via eGovernment portals, or even via large-public search engines. Procedural question-answering systems are of much interest to query legislation, court decisions, guidelines, procedures, etc. In this paper, we present a typology of how-questions asked on the web. Then, we explore facets of procedural texts : their typology and general prototypical structures. We finally present our strategy for answering procedural questions using the notion of questionability of a procedural text.

Keywords. Procedural Questions, Procedural Texts, Question-answering Systems, eGovernment Services

1. Introduction

Procedural questions, sometimes called 'How-questions', are questions whose induced response is typically a fragment, more or less large, of a procedure, i.e., a set of coherent instructions designed to reach a goal. Procedural questions form a large subset of questions typically introduced by 'How'.

Recent statistics elaborated from queries to Web search engines show that procedural questions is the second largest set of queries after factoid questions [10]. Procedural question-answering systems are of much interest both to the large public via the Web, and to more technical staff, for example to query large textual databases dedicated to various types of procedures.

Such systems are of particular interest eGovernment services since they are increasingly growing. For example, the online income tax application form was so successful in France that the government extended the deadline for internet users only. Governments are today aware of the significant benefits that can be realized by migrating traditionally face-to face services to the Internet. Administrative procedures are thus accessible via portals which are now emerging as the new e-Government single points of access for citizens and businesses [1]

[1]Correspondence to: Farida Aouladomar, IRIT, 31062 Toulouse Cedex, France. Tel.: +33 61 55 74 34; E-mail: aouladom@irit.fr.

The work presented here is a generic study designed to answer general procedural questions on the web. The answers to "how" questions can be found in procedural texts. Under the heading of procedural texts, there is a quite large diversity of texts (receipes, maintenance manuals, advice texts, etc.) [2] notes the variability of judgements in procedural text categorization, depending on the text main objectives and style. Procedural texts represent a substantial part amoung legal texts. They take the form of guidelines, circular letters, legislative texts and instructions of governmental bodies, etc.

Procedural texts share general common structures but each type of text has its own characteristics. Administrative texts are, for example, richer in conditions and contain almost no picture. Whereas assembly notices are richer in arguments, pictures, etc.

Procedural texts explain how to realize a certain goal by means of actions which are at least partially temporally organized. A procedural text can indeed be a simple, ordered list of instructions to reach a goal, but it can also be less linear, outlining different ways to realize something, with arguments, advices, conditions, hypothesis, preferences. It also often contains a number of recommendations, warnings, and comments of various sorts.

Procedural texts adhere more or less to a number of structural criteria, which may depend on the author's writing abilities and on traditions associated with a given domain. The administrative and legal domains have for instance their own writing techniques and layout. Typographical conventions such as the use of hyphens, bullets or other forms of numbering in front of each of the enumerated instructions are often imposed to writers via e.g. style files. The same is observed for a number of editing recommendations: language level, size of sentences, pronominal references, etc.

From a methodological point of view, our approach is mainly based on a corpus-based analysis, whose aim is to extract the general structure of procedural texts that is tailored for question-answering. Our corpus contains various types of procedural texts including texts as different as receipes, administrative procedures, maintenance/assembly notices, advices texts, rules, etc. In our corpus, we focused on procedural administrative texts for eGovernment documents. To get rights, citizens must often fill in forms, specifying various information. The requested information reflects the current status of the legislation and may be quite complex [7]. Application forms include, therefore, notices that help the applicant to fill them out.

In this paper, we first propose a general typology of procedural questions. In section 2, we develop an analysis of procedural texts, from the point of view of their discursive structure. In section 3, we show and evaluate the adequacy of this analysis for answering How-questions. For that purpose, we introduce different notions, among which the notion of the $questionability$ of a text, i.e. its ability to be used to answer How-questions. Finally, in section 4, we sketch out a few procedures to retrieve relevant sets of instructions from How-questions .

2. A Typology of How-Questions

Let us now investigate the structure of procedural questions. Besides an introspective analysis, the work reported below is largely based on corpora studies. We considered in particular FAQ which abound in procedural questions, questions in the TREC and AnswerBus frameworks, and quite comprehensive inventories built from queries submitted to search engines over the past month(s) composed of keywords, found at:

http://inventory.overture.com/d/searchinventory/suggestion/?mkt=fr. We constructed our procedural question corpora from different domains : legal domain, health, education, tourism, social behavior (savoir-faire and others), computer sciences, maintenance.

In this section we first identify the nature of procedural questions, since they cannot just be identified by the interrogative pronoun *how* that introduces them.

A large number of procedural questions are introduced by the interrogative pronoun *Comment* in French (How in English, Wie in German, Cómo in Spanish, etc.). However, $Comment$, similarly to How, has several uses which are not all related to procedural questions : (1) **the boolean How**: *how are you ?*, (2) **the nominal How**: *comment dit-on maison en espagnol ? (how do you say 'house' in Spanish?)*, (3) **the causality How**: *how did John die?*, (4) **the instrumental or manner How**: *how is couscous eaten in Morocco?, response: by hand* (5) **the choice list How**: *how can I pay my air ticket?, response credit card, cash, etc...*, (6) **the instructional How**: *how to change my car wheel?*

Only the last type of How questions in this analysis is typical of procedural How questions. The instrumental, manner and choice list 'How-questions' may also contain some forms of procedures.

Procedural questions are often introduced by How, but there are several other forms that we survey below:

- Forms in **'Que + Faire'** (gloss: what to do to...), as in *what should I do to get a visa for India ?* or forms with 'Quel + ETRE + proposition' (which/what + BE + prop) as in *what are the steps to follow to get a visa for India?* . We sometimes find directly constructions using the noun *procedure* to express procedural queries as in *What is the procedure to obtain the french nationality?*.
- **Elliptical use of How**: abound in cases where just two or three keywords are used instead of a full, well-formed, natural language sentence: *find lawyer, lodge complaint*. The verb or the deverbal used in the query allows for the identification of the type of question. More complex elliptical forms encountered in our corpora include e.g. the need of inference to identify a goal from a problematic situation as in: *passport lost* which must be reformulated as *how to replace my passport?*
- Questions in **is it possible to, can I, etc. + VP**, as in *is it possible to get the dual nationality?*, have a direct response which is a priori just yes or no, but, in the case of a cooperative response, which is our perspective, the response is often a set of instructions, answering the question viewed as a goal to reach.

In our system, query processing is based on the QRISTAL system (developed by Synapse Toulouse : http://www.qristal.fr), which is not perfect, but it allows us to get the distinctions presented above and to construct an adequate representation.

3. An Analysis of the Structure of Procedural Texts

In this section, we introduce our analysis of the discursive structure of procedural texts, with the view (in terms of topics and granularity) of answering how-questions.

3.1. A Discursive Analysis of Procedural Texts

Here is, represented by means of a grammar, the structure we have elaborated for procedural texts. The structures reported below correspond essentially to the organization of the informational contents. Elements concerning the layout (e.g. textual organizers such as: titles, enumerations, etc.), and linguistic marks of various sorts are used as triggers or delimiters to implement this grammar.

In what follows, parentheses express optionality, braces are used when an element is compulsory but not always expressed in words, + iteration, / is an or, the comma is just a separator with no temporal connotation a priori, and the operator $<$ indicates a preferred precedence (i.e. the elements usually appear following the elements order given in the grammar nodes). Each symbol corresponds to an XML-tag, allowing us to annotate procedural texts in order to extract relevant informations for the response formulation.

The top node is termed **Text**:
Text $\rightarrow$ **title, (summary), (warning), (pre-requisites), (picture)+** $<$ **objective.**

summary $\rightarrow$ **title+.**
Summary describes the global organisation of the procedure, it may be useful when procedures are complex (summary can be a set of hyper-links, often pointing to titles).

warning $\rightarrow$ **text , (picture)+, (pre-requisites)**.
Warnings represent global precautions or preventions associated with actions or objectives (e.g. make sure you get all the required documents prior to any naturalization application): they may have a complex rhetorical and modal structure, and should be studied in more depth. At the moment, we consider a warning just as a text, which sounds sufficient in most question-answering (QA) situations. Warnings are in fact of interest for answering $Why?$ questions.

pre-requisites $\rightarrow$ **list of objects, instruction sequences.**
Pre-requisites describe all kinds of equipments needed to realize the action (e.g. the different documents needed to fill in forms) and preparatory actions.

picture describes a sequence of charts and/or schemas of various sorts. They often interact with instructions by e.g. making them more clear. Analyzing this phenomena is outside the scope of this paper.

The objective level is the basis structure of procedural texts. It is often complex since it may contain embedded objective. Each group of instruction sequences is designed around a specific goal.
objective $\rightarrow$ **{goal}** $<$ **instruction sequences+ / objective.**

Instruction sequences is structured as follows:
instruction sequences $\rightarrow$ **instseq** $<$ **{connector}**$<$ **instruction sequences / instseq**.

Instseq is then of one of four main types below:
instseq $\rightarrow$**, imperative linear sequence / optional sequence / alternative se-**

quence / imperative co-temporal sequence.

Each type of instruction sequence is defined as follows:

imperative linear sequence → instruction < {temporal mark}, instseq/ instruction. (e.g. *En ce qui concerne votre adresse à l'étranger, indiquez le nom de la rue et le numéro s'il en existe, et ajouter votre numéro de boîte postale si le courrier n'est pas distribué à domicile dans la localité où vous résidez. Précisez enfin le nom du pays.*) (gloss : *Concerning your address abroad, mention the street name and the number if it exists and add your postal box number if mails aren't delivered at the place where you live. Finally, specify the name of the country.*) An imperative linear sequence is the kind of most common instruction sequence in procedural texts. It can be composed of one or several instructions.

optional sequence → optionality expression, imperative linear sequence. (e.g. *Si vous préférez joindre les copies des documents requis, présentez les originaux le jour de votre rendez vous au service des visas*) (gloss : *if you prefer enclosing copies of requested documents, bring originals the day you come to the visa services.*)

alternative sequence → (conditional expression), imperative linear sequence, (alternative-opposition mark) < instseq / (conditional expression, instseq)+. (e.g. *pour expédier votre demande, vous pouvez soit l'adresser directement par la poste, de préférence en recommandé, à la mairie de la commune où vous désirez vous inscrire, soit la confier au consulat, qui l'acheminera par la valise diplomatique, à l'adresse de la mairie.* (gloss : *to send your application, you can either post it, preferably in certified delivery, to the town council where you want to register, or give it to the consulate which will send it to the town council using the diplomatic bag*).

imperative co-temporal sequence → imperative linear sequence < co-temporal mark < imperative co-temporal sequence / imperative linear sequence.
A co-temporal sequence relates instructions which must be realized at the same time, or more generally non-sequentially (e.g. *Pour bénéficier des services de l'ANPE, vous devez vous inscrire à votre agence locale, et parallèlement faire une demande d'indemnisation aux Assedic*) (gloss : *to pretend to ANPE services, you should registrer to your local agency, and at the same time, make a help request to the Assedic*).

Finally, Instruction is the lowest level and has the following structure, with recursion on objective:

instruction → (iterative expression), action, (goal), (reference)+, (argument), (picture)+, (warning).

At this level, it is most important to note reference phenomena of various sorts, pointing to already described instructions, to instructions to be described later, or to external data via hyper-links (e.g. *see application form 256...*). Let us also note that we have observed almost no restatements (other ways to describe an instruction if it is difficult to understand) and no summaries synthesizing instructions (not to be confused with the generic summary of an objective, which is a kind of table of contents).

3.2. Marks for Instruction Localization

The different types of marks that allow for the identification of the different elements presented in the grammar above are presented in [4]. In this section, we present the main marks used to define the boundaries of each instruction or set of instructions.

Typographic criterion: the next point is to isolate basic instructions, called 'instructions' in the above grammar, and within these instructions the 'action' itself. The problem is essentially to identify simple marks which are delimiters of the beginning and the end of an instruction. Inter-instruction marks organize instructions. They are in general quite simple to identify in procedural texts [11]. The beginning of an instruction is often the start of a sentence which can be introduced by various typographic marks proper to enumerations (intended lines, bullets etc.) [8]. These correspond also in general to an instruction. The end of an instruction is either a punctuation mark, usually the dot, sometimes the semicolon or the comma, or typographic marks introducing the next instruction. Table 1 summarizes our observations on the use of layout to easily identify the instructions within a procedural text. Communication category includes administrative texts, advice texts, etc.

Table 1. (1) percentage of instructions or set of instructions introduced by typographic marks such as hyphens, bullets and other numbering forms, line breaks. (2) number of instructions considered in our sample.

Domains	(1)	(2)
maintenance, assembly	78%	279
receipes	89%	151
communication	63%	206
average	77%	636

Semantic criterion: within an instruction, the action is in general organized around the action verb and its arguments. Goals, references, manners, limits, are all adjuncts which appear in various orders. Goals contain specific verbs while manners are often nominal. Using the same corpora as above, Table 2 shows the following verb distribution. It is elaborated with the TROPES software.

Table 2. (1) factive verb, (2) stative verb, (3) declarative verb, (4) performative verb.

Domains	(1)	(2)	(3)	(4)
maintenance/assembly	65%	23%	12%	
receipes	85%	13%	2%	
procedural QA pairs	67%	11%	22%	
communication	52%	26%	22%	
average	67%	18%	15%	
non-procedural texts	41%	35%	23%	1%

As can be noted, procedural texts have a much higher rate of factive verbs, and much less stative verbs. declarative verbs are about the same as in other types of texts. This

verb type discrimination criterion is not precise enough for procedural text categorization, in particular to discriminate communication procedural texts from non-procedural texts.

Morphological criterion: another criterion is the morphology of the verbs encountered in procedural text. Instruction verbs are usually at the imperative, infinitive and or gerundive form. The more these forms are found in texts the more procedural these texts are. Table 3 presents the results obtained over a large list of verbs for procedural text and non-procedural texts.

Table 3. (1) number of verbs in texts, (2) number of verbs in imperative/infi nitive/gerundive form, (3) percentage w.r.t. the total number of verbs.

Texts	(1)	(2)	(3)
maintenance/assembly	826	523	63%
receipes	434	386	89%
procedural QA pairs	495	244	49%
communication	315	156	49%
total average for procedural texts	2070	1309	63%
non-procedural texts	776	200	26%

Our goal is not just to use an instruction to respond to a How-question. It is often necessary to consider the level of the 'objective' or higher, where quite generic goals, those frequently encountered in How-questions, are found. 'objectif' are in general delimited by the expression of goals, which may have various forms, and by typographic elements (e.g. starting a new paragraph). Goals may be titles, well-identified from a typographic point of view, or they may have the form of a proposition introduced by a causal mark: *to vote by proxy, ...*

4. Questionability of a Text

We use the term *questionability*, term due to J. Virbel [9], to express the ability or the relevance of any text, in our case found on the Web, to respond to How-questions. The primary goal is to have criteria to identify texts which are procedural among those obtained from a search engine. The second goal is to consider those procedural texts which are the most appropriate for responding to procedural questions. This means to be able to compare texts clearly identified as procedural texts, in terms of their level of detail, informativity, readability, conciseness, illustrations, number of links to other pages or to other parts in that same text, prevention on actions, etc.

The evaluation of the questionability of a text can be made a priori, independently from any particular query, or in relation with a query since some goals may be easier to identify than others in given texts. In this section, we establish a compromise between these two views which are of much interest. For the moment, a response to a how-question is found in a single text, using relevance, clarity, and informativity criteria. In a second stage, it would be of much interest to select a text depending on the user profile (casual user or professional) and to be able to merge or to integrate texts when they complement each other. These objectives are obviously very difficult to implement.

Let us now present the different criteria we consider to measure the questionability of a text. This measure is decomposed into two stages. The first stage aims at selecting those texts which should be a priori procedural texts. It is essentially based on surface marks to guarantee a certain efficiency. The second stage concentrates on the query, and introduces several relevance measures correlated with the query to answer. It is presented in section 6.

The first stage, called the "CATEG", selects the subset of texts returned by a search engine which should be procedural texts according to the three 'surface' criteria below; measures are all relative to text size:

- typographic forms (noted as TF) of various kinds that measure the architectural quality of the text. These forms include those given in Table 1 of section 4.2.
- morpho-syntactic marks, (noted as MSM): in procedural texts, we observed (cf. Table 3, in 4.2) that most verbs are either in the infinitive or in the imperative form, there are also marks that motivate the user to go further, such as *you must, you just have to,* followed by an action verb, or marks that indicate a task to realize: *the next stage, the next step, proceed as follows, care about, do not forget to, etc.* which abound in procedural texts,
- the presence of a large number of articulatory marks, (noted as AM): temporal, argumentative, causal marks to cite the most important ones (section 4.2).

Since it is quite difficult to assign relative weights to each of these three criteria, we consider they have an equivalent weight in the selection of procedural texts. Each counts for a third of the decision. Given a set of n texts, we evaluate for each text TF, MSM and AM. For example, TF_i is the ratio: number of typographic forms divided by the size of the text in number of words in the text i. Then, for each criterion, the average frequency is computed:
$TF_{average} = (\Sigma_{i=1,n} TF_i)/n$,
and similarly for the other two criteria. We can now define the CATEG for text i w.r.t. the set of texts considered:
$CATEG_i = TF_i/TF_{average} + MSM_i/MSM_{average} + AM_i/AM_{average}$.

The second stage, the "QUEST", investigates in more depth the questionability of the text. The objective is to evaluate the number of areas which can potentially match with How-questions. This is carried out by identifying those areas in the text on which the matching with questions should potentially be realized. Via our corpora analysis, it turns out that those areas are essentially:

- the number of titles identified, under objectives and in the summary (noted as TIT),
- the presence of a large number of action verbs (noted as AV), (cf. Table 2, in 4.2),
- the number of goals identified, (1) associated with instruction sequences or (2) within basic sequences, associated with the action to realize, (noted as GOA), and, finally
- manners found in instructions (noted as MAN).

Different linguistic marks allow for the identification of the goals and manner : the causal and manner connectors(e.g. in order to, so, by + gerundive verb, with, etc.).

Similarly as above, we can define the QUEST rate for a given text i in a collection of n texts. $QUEST_i = TIT_i/TIT_{average} + AV_i/AV_{average} + GOA_i/GOA_{average} + MAN_i/MAN_{average}$.

We can then compute an estimate of the overall questionability of a text i in a collection of n texts as follows: $CATEG_{average} = (\Sigma_{i=1,n} CATEG_i)/n$,
$QUEST_{average} = (\Sigma_{i=1,n} NN_i)/n$,
$questionability_i =$
$CATEG_i/CATEG_{average} + QUEST_i/QUEST_{average}$.

5. Responding to How-questions

The main aim of this project is to adequately and cooperatively respond to How-questions. An accurate and relevant analysis of the structure of How-questions, of procedural texts and of the notion of questionability establishes a basis for associating a response with a query which is as adequate as possible. Within our present perspective, responding to how-questions involves the following tasks:

- selecting the procedural texts which have the best questionability rate. Since, at this level, the matching with the query has not yet been done, we keep the 20 best texts based on the metrics given above,
- matching the question body with 'questionable zones' of procedural texts, hierarchically organized as: titles, goals, manners, and defining the best match,
- extracting the relevant portion of the text and returning it to the user in a user-friendly way.

The last step consists in selecting the appropriate text fragment that responds the question. So far, our strategy is quite simple, and we have the following main situations:

- If the question matches with the title of the whole document, then the document is selected as a whole,
- If the question matches with the title or the goal of an instruction sequence, as defined in the grammar, then that whole sequence is selected. This is however a general rule which suffers some exceptions. In particular, for alternative sequences, it may be useful to select a larger fragment of the text.
- If the question matches with a goal within an instruction, then this instruction is returned to the user.

The adequacy of this rough strategy remains to be evaluated in depth. Since it is not easy to predict when a larger text fragment will be necessary, our strategy is to return a window that displays a priori the selected portion, however, the user can scroll it up or down to get a larger or a nearby text portion. Besides the response, in case of an indirect match (e.g. using more generic terms or synonyms), an explanation must be provided so that the user understand why he gets such as response. The explanation follows the template philosophy presented in WebCoop [6], outlining the terms that have been changed and why.

6. Perspectives

In this paper, we presented the general structure of procedural texts. We also investigated the structure of How-Questions, outlining those which really induce responses under the

form of sets of instructions. We then showed how a procedural text can be characterized using relatively external and simple criteria. Finally, we briefly presented how to characterize the questionability of a text, and how the response retrieval mechanisms can be constructed from this notion.

This work is still very experimental, it raises many questions. Our work needs to be deepened along many lines, including response accuracy (quality and scope), response generation, and the development of more elaborated evaluation methods, much more complex than e.g. the methods used in TREC for factoid questions.

Works on procedural question-answering systems has many applications in the legal domains and must be deepened and adapted to them. This work can support general public "how do I" questions for e-governments (administrative procedures) or also "how" questions dealing with court files, and court decisions for instance. Other legal-field documents such as legislation texts which have specific structure with specific linguistic marks and lay-out, must be investigated. Question answering systems in the legal field are both useful for large public and practicioners of the domain.

References

[1] Accenture report. eGovernment Leadership : Rhetoric vs Reality - Closing the Gap. http://www.accenture.com/xdoc/en/industries/government/2001FullReport.pdf, 2001.

[2] Adam, J.M.. *Types de Textes ou genres de Discours ? Comment Classer les Textes qui Disent De et Comment Faire*, Langages, 141, pp. 10-27, 2001.

[3] Allen, J., *Towards a General Theory of Action and Time*, Artifi cial Intelligence, 23:123-154, 1984.

[4] Aouladomar, F., Saint-Dizier, P., *An Exploration of the Diversity of Natural Argumentation in Instructional Texts*, 5th International Workshop on Computational Models of Natural Argument, IJCAI, Edinburgh, 2005.

[5] Aouladomar, F., *A Preliminary Analysis of the Discursive and Rhetorical Structure of Procedural Texts*, Symposium on the Exploration and Modelling of Meaning, Biarritz, 2005.

[6] Benamara, F., Saint-Dizier, P., *Advanced Relaxation for Cooperative Question Answering*, in: New Directions in Question Answering, in Mark T. Maybury, (ed), AAAI/MIT Press, 2004.

[7] The GIST Project, *Generating InStructional Text*, http://tcc.itc.it/history/projects/gist.html, 1994-1996.

[8] Luc, C., Mojahid, M., Virbel, J., Garcia-Debanc, C., Pery-Woodley, M-P., *A Linguistic Approach to Some Parameters of Layout: A study of enumerations*, In R. Power and D. Scott (Eds.), Using Layout for the Generation, Understanding or Retrieval of Documents, AAAI 1999 Fall Symposium, pp. 20-29, 1999.

[9] de Mattos Pimenta Parente, M-A., Steffen Holderbaum, C., Virbel J., Nespoulous, J-L., *Text Questionability as a predictor of story recall*, Thirteen Annual Meeting of the Society for Text Understanding, Madrid, 2003.

[10] De Rijke, M., *Question Answering: What's Next?*, the Sixth International Workshop on Computational Semantics, Tilburg, 2005.

[11] Takechi, M., Tokunaga, T., Matsumoto, Y., Tanaka, H., *Feature Selection in Categorizing Procedural Expressions*, The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL2003), pp.49-56, 2003.