A Computational Model for Generating Referring Expressions in a Multilingual Application Domain

Elena Not
IRST
Loc. Pantè, I-38050 Povo - Trento, Italy
not@irst.itc.it

Abstract

In this paper we analyse the problem of generating referring expressions in a multilingual generation system that produces instructions on how to fill out pension forms. The model we propose is an implementation of the theoretical investigations of Martin and is based on a clear representation of the knowledge sources and choices that contribute to the identification of the most appropriate linguistic expressions. To cope effectively with pronominalization we propose to augment the Centering Model with mechanisms exploiting the discourse structure. At every stage of the referring expressions generation process issues raised by multilinguality are considered and dealt with by means of rules customized with respect to the output language.

1 Introduction

An automatic generation system that is to produce good quality texts has to include effective algorithms for choosing the linguistic expressions referring to the domain entities. The expressions have to allow the reader to easily identify the referred objects, avoiding ambiguities and unwanted implications. They have to conform to the expectations of the reader according to his evolving flow of attention and they have to contribute to the cohesion of the text.

In this paper, we describe a component, developed inside the GIST project[1], building referring expressions for automatically generated multilingual (English, German, Italian) instructions in the pension domain. The overall decision making mechanism for the referring expressions choices is based on the theoretical investigations of Martin (Martin, 1992), for which we propose a possible implementation. The implemented model proved to be particularly suitable to work in a multilingual domain. For the generation of pronouns we define an extension of the Centering Model exploiting the contextual information provided by the rhetorical structure of discourse (Not and Zancanaro, 1996).

At every stage of the referring expressions generation process issues raised by multilinguality are considered and dealt with by means of rules customized with respect to the language. In section 2 we first present the results of observations made on the corpus texts with the aim of identifying the typical referring expressions occurring in our domain. Section 3 details the solutions implemented in the GIST system. Specifications for the implementation are given in terms of data structures and required algorithms.

2 Referring expressions in multilingual pension forms

Our work on the specification for the referring expressions component started from an analysis of the collected multilingual (English, German, Italian) corpus of texts containing instructions on how to fill out pension forms. From this study, a general typology of entities referred to in the domain emerged together with an indication of how such entities are typically expressed in the different languages (see figure 1). The classification includes:

Specific entities. These are extensional entities: individuals or collections of individuals (plurals). In KL-One knowledge representation languages they are represented as instances in the A-box.

Generic entities. These entities are intensional descriptions of classes of individuals and are often mentioned in administrative doc-

[1] The GIST consortium includes academic and industrial partners -IRST (Trento, Italy), ITRI (University of Brighton, Great Britain), ÖFAI (Vienna, Austria), Quinary (Milano, Italy), Universidade Complutense de Madrid (Spain)- as well as two user groups collaborating actively to the specification and evaluation of the system -INPS (the Italian National Security Service) and the Autonome Province of Bolzano.

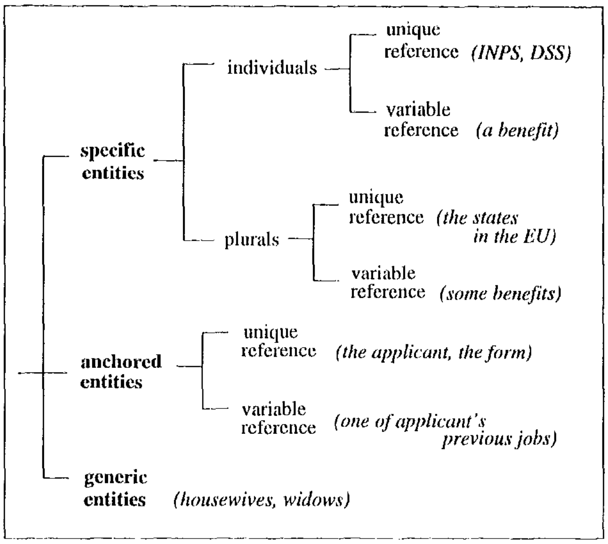
specific entities
individuals
unique reference (INPS, DSS)
variable reference (a benefit)
plurals
unique reference (the states in the EU)
variable reference (some benefits)
anchored entities
unique reference (the applicant, the form)
variable reference (one of applicant's previous jobs)
generic entities (housewives, widows)

Figure 1: Types of entities referred

uments, since the entities (persons or inanimate objects) addressed in this kind of texts are not usually specific individuals in the mind of the public administrator but rather all the individuals that belong to a certain class, as in the following example:

(1) Married women should send their marriage certificate.

In KL-One knowledge representation languages generic entities are represented as concepts in the T-box.

Anchored entities. These are entities that, although generic in nature, can be interpreted as specific when they are considered in the specific communicative situation in which the actual applicant reads the instructions to complete the pension form he has in his hands. Consider for example the following text: "The applicant has to provide all the requested information". In this situation, the specific person who is reading the form instantiates the generic applicant. All the entities directly related to the applicant or to the form itself can be considered anchored, as for example: the applicant's name, the applicant's spouse, any applicant's previous job, section 3 of the form, ... The plausibility of this anchoring operation is confirmed by the fact that the linguistic realization choices made for anchored entities (definite forms, singular indefinite forms, ...) resemble very much the linguistic choices made for specific entities.

Further investigations on the corpus texts have been conducted to identify language-dependent referring phenomena and general heuristics for the choice of the most appropriate linguistic realization. In general, we found that language style has great influence on the realization choices. When an informal style is used (like in most English documents and in some recent Italian/German forms) the personal distance between the interlocutors (the citizen and the public institution) is reduced using direct references to interlocutors, by means of personal pronouns ("you", "we"). When the language is more formal, impersonal forms or indirect references are preferred ("the applicant", "INPS", "DSS").

Apart from style differences, there do exist also differences in realization that depend on the output language. For example, in administrative forms, in full sentences, for entities anchored to the reader English and German typically use possessive noun phrases (like "your spouse") whereas Italian prefers simple definite forms (e.g. "il coniuge" [the spouse]).

3 The adopted approach

The linguistic expressions that refer in the text to the domain entities have to fulfill several properties:

- they must allow the non-ambiguous identification of the entities[2];
- they should avoid redundancies that could hamper fluency;
- they should contribute to the cohesion of the text by signaling semantic links between portions of text;
- they should conform to the formality and politeness requirements imposed to the output texts.

When we choose to realize a referring expression with an *anaphora* we fulfill a double function: we introduce some form of economy for the reference, avoiding the repetition of a long linguistic expression, and we enhance the coherence of the text since we signal meaning relations (*cohesive ties*) between portions of the discourse.

The choice of the correct referring expression depends on two major factors:

(A) the cohesive ties that we want to signal to improve the cohesion of the text;

(B) the semantic features that allow the identification of the object in the domain (distinguishing semantic features).

Another relevant factor is the pragmatic setting of the discourse (formality and politeness).

To decide on (A), data structures are maintained that keep track of the evolving textual context (discourse structure and focus history) and record the socio-cultural background of the reader

[2] In some genres the use of ambiguous references may be possible or desirable, for example in jokes, but in administrative genre clearness and unambiguity are the primary goals.

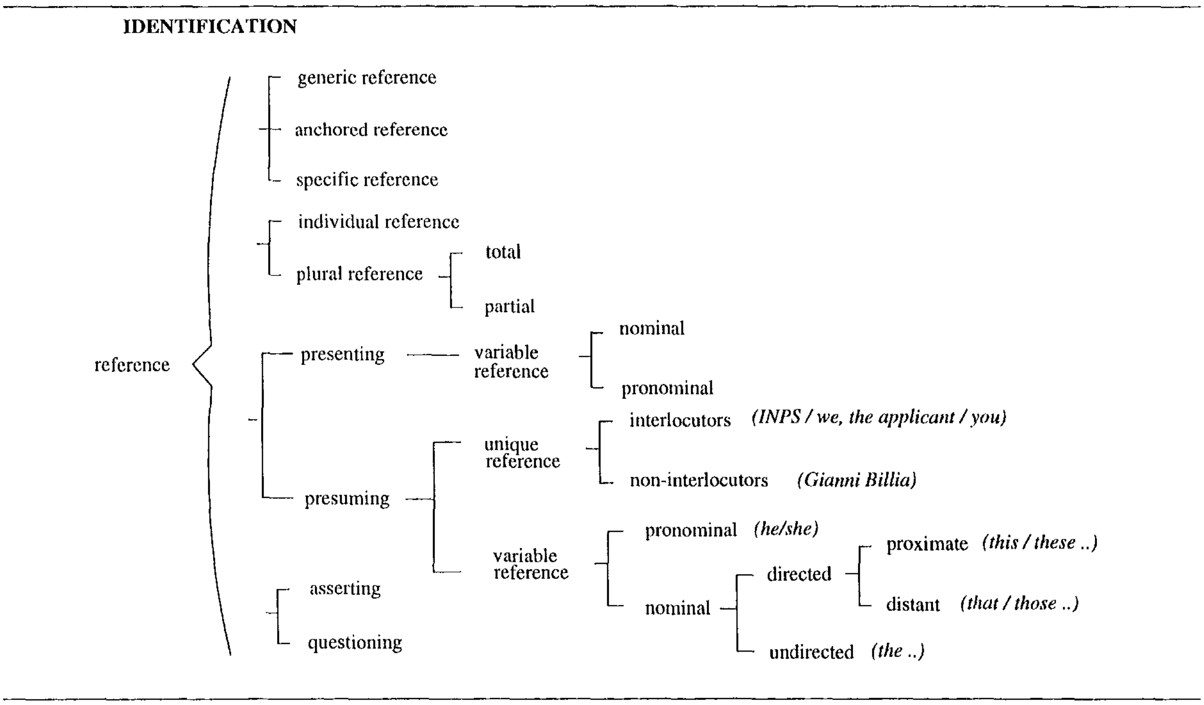

Figure 2: How semantic features combine to identify the entity in the context

(user model). Inquiries on these data structures are performed to verify whether the identity of the entity can be recovered from the context or whether there exist semantic relations with other cited entities that are worth being signaled (e.g. comparative relations).

Once the ties have been determined, the distinguishing semantic features are identified. These semantic features depend on the entity type – whether generic, anchored or specific – and on the relationships between the entity and the context – whether the entity is new with respect to the context (*presenting*) or its identity can be recovered (*presuming*). Figure 2 illustrates a fine grained distinction of semantic features whose combination specify how a referring expression can be built. This network of choices is an adaptation to the GIST application domain of the results presented in (Martin, 1992).

The *total/partial* opposition is used to distinguish references to sets of elements from references to portions of sets. The linguistic form of the expression also varies according to the type of speech act that is to be realized, and this justifies the *asserting/questioning* distinction.

Entities may be presented as new in the discourse context through references composed by a nominal expression or a pronoun (*presenting*).

A presupposed element (*presuming*) may belong to the cultural/social context, and therefore be described with a *unique* reference, or it may belong to the textual context. The *presuming-variable* option corresponds to a textual anaphora. In this case a pronoun or a definite expression can be used. In our system, pronominalization is decided according to new rules extending the Centering Model, as explained in the following section 3.3. Definite expressions are built selecting the appropriate determiner (the, this, that...) and the information (head, modifiers) to put in the noun phrase. This latter information is determined through the algorithm explained in section 3.2.

3.1 The global algorithm

The submodule for the generation of referring expressions is called during the final stage of the text planning process, when the so called microplanning (or sentence planning) takes place (Not and Pianta, 1995). The global algorithm implemented has been derived from the network of choices presented above, as emerging from the corpus analysis. The formal approach adopted proved to be particularly suitable to cope with multilinguality issues, since the tests performed at the various choice points can be easily customized according to the output language. The algorithm is activated on each object entity to be referred and accesses the following available contextual information:

Background - the cultural and social context of the reader. At present this is represented by a list of all the entities the reader is supposed to know (e.g. the Department for Social Security, the anchored entities);

RT - the rhetorical tree, specifying how the selected content units will be organized in the final text and which are the semantic relations between text spans that will be signaled to enhance the coherence;

AlreadyMentioned - the history of already mentioned entities;

StylePars - the parameters that define the style of the output text;

FocusState - the state of the attention of the reader, organized as detailed in section 3.3.

To model the rhetorical structure of discourse we consider the Rhetorical Structure Theory as developed in (Mann and Thompson, 1987). According to this theory, each text can be seen as a sequence of clauses linked together by (semantic) relations. These relations may be grammatically, lexically or graphically signaled. About 20 such relations have been identified by (Mann and Thompson, 1987), e.g. ELABORATION, which occurs when one clause provides more details for a topic presented in the previous clause, CONTRAST which links two clauses describing similar situations differing in few respects, and so on.

Here follows a sketch of the global algorithm implemented (Not, 1995). To make the reading easier, labels in italics have been introduced to identify the steps of the algorithm corresponding to the main choice points in figure 2.

```
Preliminary step:
• (For English) if e is an anchored entity treat it as
    if it was a specific entity in Background
• (For Italian and German) if e is an anchored entity
    inside a concept description
      then treat it as a presenting of a generic entity
        with a nominal expression (goto presenting-
        nominal-generic)
      else treat it as if it was a specific entity in Background
In case:
• e is referred to in a title and is anchored to the
    reader
    (English, German) if *formality* = informal
        then use a noun phrase with the possessive
        adjective "your"
        else use a bare noun phrase
    (Italian) use a bare noun phrase
• e is referred to in a title (but is not anchored to
    the reader) or in a label
    use a bare noun phrase (singular or plural according to the number of e)
• e ∈ AlreadyMentioned ∪ Background
    then [presuming]: if e ∈ Background
      then [unique]: if e is-a interlocutor
        then [interlocutor]: if *formality* = informal
          then use a pronoun
          else use a proper noun (if it exists) or
          a definite description
        else [non-interlocutor]:
          (English, German) if *formality* =
          informal and e is anchored to the
          reader
            then use a noun phrase with the possessive adjective "your"
            else use a proper name or a definite
            description
          (Italian) use a proper name or a definite description
      else [variable]: attempt pronominalization
        using the algorithm described in section
        3.3 accessing FocusState and RT. If e is
        pronominalizable
          then [pronominal]: use a pronoun
          else [nominal]: build an anaphoric expression. Test FocusState to identify
          the most appropriate determiner for
          the noun phrase. Compute the head
          and the modifiers using the algorithm
          described in section 3.2.
    else [presenting]: if e stands for a generic person
      (collection of persons) without any specific
      property
      then [pronominal]: use an indefinite pronoun
      else [nominal]: build a noun phrase, choosing the appropriate linguistic form
        If e is a:
        - specific entity, build an indefinite singular description or an indefinite plural description according to the number of e
        - generic entity, in case:
            • e is a concept whose meaning is being defined by syntesis, use the bare
            singular term
            • e is a concept being defined
            through a listing of its components,
            use a definite singular noun phrase
            • e appears in a list inside a concept
            definition,
            (German, Italian) use a bare singular or bare plural noun phrase
            (English) use a definite singular or
            definite plural noun phrase
            • e is in a question, use a singular
            indefinite noun phrase
            • e is used in procedural descriptions,
            (Italian, German) use a definite plural description.
            (English) use a bare plural.
```

3.2 Generating nominal expressions

In this section we focus on the choice of the head and the modifiers for noun phrases. (Dale and Reiter, 1995) contains the following list of requirements for a referring expression to obey to Grice's Maxims of conversational implicature:

1. The referring expression should not include unnecessary information (the Maxim of Quantity).
2. The referring expression should only specify properties that have some discriminatory power (the Maxim of Relevance).
3. The referring expression should be short (the Maxim of Brevity).
4. The referring expression should use basic-level and other lexically preferred classes whenever possible (Lexical Preference).

Requirement (4) suggests that the head of the noun phrase should be chosen among terms of common use or, more in general, among terms that the user is likely to know. In our domain, however, often technical terms can not be avoided since the precise type of document or legal requirement have to be specified. Therefore, for the choice of the head of non-anaphoric expressions the GIST system adopts the strategy of using the most specific superconcept of the entity that has a meaningful lexical item associated (e.g. the specific term "decree absolute" is used instead of the more basic term "certificate").

Requirements (1),..,(3) suggest that the modifiers in the noun phrase should not introduce unnecessary information that can hamper the text fluency and yield false implications. The task of selecting the correct modifiers for a non-anaphoric expression is not an easy task, since in the Knowledge Base attributive and distinguishing (restrictive) properties are mixed. In GIST, the semantic relations that are relevant in the definition of distinguishing descriptions have been identified through an accurate domain analysis. For example, we have chosen relations like has-partnership, owned-by or attribute-of, characterizing distinguishing descriptions like "the applicant's spouse" or "the applicant's estate".

When an anaphora occurs but a pronoun can not be used, a nominal anaphoric expression is built. The head and the modifiers included in the noun phrase have to allow the identification of the entity among all the ones active in the reader's attention (*potential distractors*). In GIST we adopt an algorithm which is a simplified variation of the one Dale and Reiter call the "Incremental Algorithm" (Dale and Reiter, 1995): whenever a new nominal anaphoric expression has to be built, discriminant modifiers are added to the expression until the set of the potential distractors (*contrast set*) is reduced to an empty set.

3.3 Generating pronouns

For the generation of pronouns an extension to the Centering Model (Grosz et al., 1995) has been defined that captures how the rhetorical evolution of the discourse influences the flow of attention of the reader. The choice of this solution has emerged from the observation that anaphora plays two roles in the discourse: it is not sufficient that a pronoun identifies unambiguously its referent but it has to reinforce the coherence of the text as well, supporting the user's expectations.

In the Centering Model for each utterance U_n a list of *forward looking centers*, $Cf(U_n)$, made up of all the entities realized in the utterance, is associated. This list is ordered according to the likelihood for the elements of being the primary focus of the following discourse. The first element in the list is called the *preferred center*, $Cp(U_n)$. Among the centers another significant entity is identified: the *backward looking center*, $Cb(U_n)$. This represents the primary focus of U_n and links the current sentence with the previous discourse.

The basic constraint on center realization is formulated in the following rules:

RULE 1 : If any element of $Cf(U_n)$ is realized by a pronoun in U_{n+1} then the $Cb(U_{n+1})$ must be realized by a pronoun also. (Grosz et al., 1995)

RULE 1' : If an element in $Cf(U_{n+1})$ is coreferent with $Cp(U_n)$ then it can be pronominalized. (Kehler, 1993)

These rules can be used to constrain pronominalization in the text generation process.

The Centering Model was first conceived for English, a language where pronouns are always made explicit. But as soon as we consider languages that allow null pronominalization (like Italian) new extensions to the original model have to be designed in order to deal with pronouns with no phonetic content. For Italian, we defined the following rule (Not and Zancanaro, 1996) which is compatible with the results of empirical research presented in (Di Eugenio, 1995):

RULE 1" : If the Cb of the current utterance ($Cb(U_{n+1})$) is the same as the Cp of the previous utterance ($Cp(U_n)$) then a null pronoun should be used. If, instead, $Cb(U_{n+1}) \neq Cp(U_n)$ and $Cb(U_{n+1}) = Cb(U_n)$ then a strong pronoun should be used.

3.3.1 The proposed extension to the Centering Model

Unfortunately, the Centering Model does not capture completely the reader's flow of attention process since it fails to give an account of the expectations raised by the role the clause plays in the discourse. For example consider the following sentences:

(2) a. If you are separated,
b. [your spouse]$_i$ should send us [this part of the form]$_j$ properly filled in.
c. [They]$_i$ should use [the enclosed envelope]$_k$.
d. e_k does not need a stamp.

According to the Centering rules it would not be possible to use a pronoun to realize e_k since the main center of utterance **d.** (the envelope) is different from the main center of utterance **c.** (the spouse). But the use of a definite noun phrase to refer back to the envelope would sound rather odd to a native speaker.

However, the rhetorical structure of the text, providing information on the semantic links between utterances, helps understanding how the content presentation progresses. Therefore, we claim that it can be used to explain exceptions to the Centering rules and used to define repairing strategies (Not and Zancanaro, 1996). The advantage of this solution is that it allows us to treat with a uniform approach different types of exceptions that in literature are solved with separated ad-hoc solutions (e.g. parallelism, empathy).

For example, in (2) above sentence **d.** is an evident ELABORATION on the envelope that appears in sentence **c.** When elaborating the description of an object the focus of attention moves onto the object itself. Therefore, the rhetorical relation that links **c.** and **d.** signals that among the elements in Cf(c) the envelope is the best candidate to be the primary focus of the following sentence **d.** This means that the rhetorical information can "project" the default ordering of the elements in the potential focus list Cf(c) onto a new order that reflects more closely the content progression.

From a computational point of view, the resulting algorithm for pronominalization can be sketched as follows. The reader's attentional state is recorded in two stacks: the *Centers History Stack* and the *Backward Centers Stack* collecting respectively the Cf and the Cb of the already produced utterances. Whenever a new utterance is processed, the corresponding Cf and Cb are pushed on the top of the two stacks. The Cf list is ranked according to the default ranking strategy:
clause theme > actor > benefic. > actee > others
possibly modified by a "projection" imposed by the rhetorical relation. Rules 1' (for English and German) and Rule 1" (for Italian) are then used to decide whether a pronoun can be used or not.

4 Conclusion

We have presented the computational model implemented in the GIST system for referring expressions generation. The model is based on a clear distinction of the various knowledge sources that come into play in the referring process and provides an implementation for Martin's theoretical investigations. An extension of the Centering Theory has been proposed to deal with pronominalization effectively, exploiting the information provided by the discourse structure on how the reader's flow of attention progresses. Issues of multilinguality are treated by customizing the selection rules according to the output language.

5 Acknowledgments

The global approach to the generation of anaphoric expressions presented in this paper, and in particular the treatment of pronominalization, has been developed together with Massimo Zancanaro, whose help I gratefully acknowledge.

References

Robert Dale and Ehud Reiter. 1995. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19:233–263.

Barbara Di Eugenio. 1995. Centering in Italian. In Ellen Prince, Aravind Joshi, and Lyn Walker, editors, *Centering in Discourse*. Oxford University Press.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), June.

Andrew Kehler. 1993. Intrasentential Constraints on Intersentential Anaphora in Centering Theory. In *Proceedings of Workshop on Centering*, University of Pennsylvania.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. In Livia Polanyi, editor, *The Structure of Discourse*. Ablex Publishing Corporation.

J. R. Martin. 1992. *English Text. System and Structure*. John Benjamins Publishing Company.

Elena Not and Emanuele Pianta. 1995. Issues of Multilinguality in the Automatic Generation of Administrative Instructional Texts. In M. Gori and G. Soda, editors, *Topics in Artificial Intelligence, Proceedings of the Fourth Congress of the Italian Association for Artificial Intelligence*, Lecture Notes in Artificial Intelligence. Springer. Also available as IRST Technical Report #9505-17, May 1995.

Elena Not and Massimo Zancanaro. 1996. Exploiting the Discourse Structure for Anaphora Generation. In *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium, DAARC96*, Lancaster University, 17-18th July.

Elena Not. 1995. Specifications for the Referring Expressions Component. Technical Report GIST LRE Project (062-09), Deliverable TSP-4, IRST, September.