Automatic Generation of Dictionary Definitions from a Computational Lexicon

Penny Labropoulou, Elena Mantzari, Harris Papageorgiou, Maria Gavrilidou

Institute for Language and Speech Processing Epidavrou & Artemidos 6, 151 25 Maroussi, Greece {penny, elena, xaris, maria}@ilsp.gr

Abstract

This paper presents an automatic Generator of dictionary definitions for concrete entities, based on information extracted from a Computational Lexicon (CL) containing semantic information. The aim of the adopted approach, combining NLG techniques with the exploitation of the formalised and systematic lexical information stored in CL, is to produce well formed dictionary definitions free from the shortcomings of traditional dictionaries. The architecture of the system is presented, focusing on the adaptation of the NLG techniques to the specific application requirements, and on the interface between the CL and the Generator. Emphasis is given on the appropriateness of the CL for the application purposes.

1. Introduction

In this paper, we present some issues concerning the construction of an automatic Generator of definitions from information stored in a Computational Lexicon (CL), namely the semantic lexicon developed in the framework of the SIMPLE project.

The aim of constructing such a system is to remedy the lack of systematicity that is often observed in definitions of traditional dictionaries intended for the human user. To accomplish this, we take advantage of the formalised, systematically and explicitly coded lexical information stored in Computational Lexica in order to generate systematic and structured definitions in a controlled natural language. The approach we have adopted for our application presents further advantages:

- the Natural Language Generation (NLG) techniques we have applied in our system lead to the generation of fluent natural language text which is familiar to the dictionary user, and
- the hierarchical organisation of information in the suggested CL contributes to the generation of well formed definitions.

The application, as presented in this paper, concerns the generation of **definitions for a monolingual medium-size general language dictionary for the average native speaker**, and focuses on definitions for concrete entities. The decisions concerning the architecture and design of the system are greatly influenced by the context of this application. Important issues to be taken into account as to the contents and structure of the target definition include the following:

- a "*medium-size dictionary*" purports to illustrate the meaning(s) of a word via a <u>comprehensible</u> and <u>concise</u> definition
- a "general language dictionary" purports to give the meaning of a word for the layman rather than to provide a full encyclopaedic or scientific definition of the object being described
- the *user* of such a dictionary is unfamiliar with the meaning of a specific word; therefore, the desired definitions aim at being <u>informative</u> via <u>valid and accurate pieces of information</u> realised in meaningful well formed text chunks.

In the following sections, after a brief presentation of relevant work conducted in the areas of NLG and Natural Language Processing (NLP), we proceed to the description of the specific application. Section 3 presents the requirements analysis for the task of dictionary definitions generation, and the methodology according to which this has been carried out. Section 4 briefly describes the structure of the CL we have used as our knowledge base, illustrating the appropriateness of our resources for the intended application. Section 5 elaborates on the system architecture, focusing on the intercommunication of the CL with the NLG system, and the adaptation of state-of-the-art NLG techniques to the requirements of an application in the domain of lexicography. Finally, Section 6 concludes with the advantages and disadvantages of the proposed system, and Section 7 gives potential future work directions.

2. Background

This work shares some common features with work done in the area of NLG systems aiming at generating descriptions of objects in the domains of encyclopaedias (Milosavljevic & Dale, 1996), museum guided tours (Mellish et al., 1998; Cox et al., 1999), and descriptions of monuments (Carenini et al., 1993), since the ultimate goal is that of describing concrete entities. However, as will be shown in the following section, the requirements of generating dictionary definitions for concrete entities differ considerably from the requirements of describing objects in the above mentioned applications.

In the area of Computational Lexicography, there is an important number of efforts purporting to extract Machine lexical information from Readable Dictionaries in order to use it in NLP applications (see, for example, (Boguraev & Briscoe, 1989; Calzolari & Picchi, 1998; Calzolari & Briscoe 1992; Copestake, 1990; Neff & McCord, 1990; Sanfilippo & Poznanski, 1992)). The effort presented in this paper, on the other hand, consists in the reverse procedure, that of extracting information stored in CL and "translating" it into fluent natural language so as to render it comprehensible to human end-users.

3. Requirements analysis

3.1. Description of the process

In the Introduction, we have identified the general framework in which we have placed our application, that of medium-size general language dictionaries, and have presented its basic requirements. In order to further specify the particular requirements of the task of defining concrete entities, we have adopted a corpus analysis approach, and, more specifically, an analysis of the content and structure of existing dictionary definitions for concrete entities.

The relevant corpus was extracted from three medium-size general language dictionaries of modern Greek (Babiniotis, 1998; Instituto Neoellinikon Spoudon, 1998; Kriaras, 1995). In order to select the entries for which the definitions would be collected for study, we have focused on those entries from our input CL to which the semantic type [Concrete_entity] or one of its subordinate types was assigned (see Figure 1). Our research was further restricted to a sample of 1,000 entries corresponding to a representative number for each semantic type.

CONCRETE_ENTITY

- LOCATION (3_D_location, Geopolitical_location, Area, Opening, Building, Artifactual_area)
- MATERIAL
- ARTIFACT (Artifactual_material, Furniture, Clothing, Container, Artwork, Instrument, Money, Vehicle, Semiotic artifact)
- FOOD (Artifact_Food, Flavouring)
- PHYSICAL_OBJECT
- Organic_object
- LIVING_ENTITY
 - o Animal (Earth_animal, Air_animal, Water_animal)
 - Human (People, Role [Ideo, Kinship, Social_status], Agent_of_temporary_activity,
 - Agent_of_persistent_activity, Profession)
 - Vegetal_entity (Plant, Flower, Fruit)
 - 0 Micro-organism
- SUBSTANCE (*Natural_substance, Substance_food, Drink* [Artifactual_drink])

Figure 1: The SIMPLE ontology for concrete entities

The analysis was carried out separately for each group of entries belonging to the same semantic type; the separate observations were then collected and the generalisations that have been made over the total sample are presented hereafter.

3.2. Corpus analysis results

As regards the *content*, dictionary definitions typically express the meaning of a word via a "genus" term, which designates a superordinate concept, and the "differentiae", i.e. those characteristics required to differentiate it from the other words located in the same level of the hierarchy. More specifically, our corpus analysis results have shown that, in the case of concrete entities, definitions consist of the following constituents:

• information which <u>classifies</u> the entity being defined into a group of similar objects,

e.g. entities belonging to the semantic type [Instrument] are being defined as belonging to one of the following groups: musical instruments / tools / medical instruments / measuring instruments etc.

- information describing the <u>properties</u> of the physical appearance, constitution or typical attributes of the entity, such as size, composition, colour, shape etc.
 e.g. entities belonging to the semantic type [Instrument] include information on the parts they are composed of (i.e. whether they contain strings, keyboard, handle, blade, etc.)
- information concerning the <u>provenance</u> of the entity,

e.g. entities of the [Fruit] semantic type include information on the trees that produce them

• information concerning the <u>functions</u> or use of the entity,

e.g. for entities of the [Furniture] semantic type, the specific use for which they are constructed is given.

Naturally, not all definitions include all these types of information. Depending on the semantic type which an entity is classified into, some of the above information may or may not be present according to the relevance to the entity's nature; for instance, information concerning the provenance of an entity is not included in definitions of entities belonging to the type [Location], but it is always included in the definitions of entities belonging to the type [Fruit].

As regards the *structure*, definitions of concrete entities in the medium-size dictionaries that we have investigated, are typically composed of a single elliptic sentence lacking the main verb. In fact, they consist of a nominal phrase, where the head is the hypernym of the defined entity, which also governs the relative clauses, prepositional phrases, participles, etc. that follow it.

Our corpus analysis has pointed to the following deficiencies in the definitions:

- lack of systematicity in the amount and types of information included in the respective definitions: for instance, the definitions of some entities of the type [Furniture] included information on the material which they are made of, while others did not; this may be due to lack of clear specifications as to which information is considered crucial for the semantic distinction of entities belonging to the same semantic type and should, therefore, be included in their definitions
- "definition" of some entries through a list of synonyms rather than through the analysis of their semantic content, as presented above
- definition of some entries through the use of highly technical terms or by reference to other entities, based on the assumption that the user of the dictionary is already familiar with them
- lack of systematicity in the lexical choice of the genus and differentiae terms within the same dictionary: for instance, entities of the type [Furniture] that serve for sitting (e.g. chair,

armchair, etc.) are defined as "furniture used to sit on", "sitting place", "seat", etc.

• lack of systematicity in the linguistic realisation of the information: for instance, the attribute size can be realised as an adjectival phrase (AP), as a nominal phrase in the genitive case (NP_gen) or as a prepositional phrase (PP).

The corpus analysis has, therefore, provided us with the application requirements, in the sense that it has indicated the content and structure of the desired definitions, and has pointed to the shortcomings that should be avoided in our application output.

4. Resources

The desired output definitions should include the four types of information identified through our corpus analysis (Section 3.2). Thus, we have proceeded to the investigation of our input CL in order to detect whether and where, if present, this information resides.

4.1. Description of the PAROLE/SIMPLE CL

Our input is a CL designed for NLP applications. More specifically, the lexicon from which the definitions will be generated, is the ILSP CL, which has been constructed according to the PAROLE/SIMPLE model for the Greek language (Gavrilidou et al., 1998; Gavrilidou et al., 2000). The CL is structured in three levels corresponding to morphological, syntactic and semantic information and is implemented in SGML format.

The PAROLE/SIMPLE Computational Lexica have been built in a harmonized way for 12 European languages, based on a common model, methodology and linguistic specifications. The PAROLE model specifies the information to be included in the CL at the morphological and syntactic level, while the SIMPLE model caters for the semantic level. For each language, 20,000 entries are coded at the morphological and syntactic level and 10,000 at the semantic level.

The contents of the semantic level of the CL constitute the knowledge base from which the main bulk of information is extracted for our purposes. The full model is described in (SIMPLE Specifications Group, 2000); in this paper, we will restrict our presentation to aspects of the model relevant to our work.

The basic unit of the PAROLE/SIMPLE model is the Morphological Unit, corresponding to a lexical item, and carrying morphological information (i.e. inflection, grammatical category classification, etc.). The Morphological Unit is linked to at least one Syntactic Unit, which bears information on the syntactic behaviour of a lexical item. Each Syntactic Unit is linked to at least one Semantic Unit, which is the basic carrier of semantic information.

Each Semantic Unit corresponds to a word sense and includes all types of information that can identify this particular word sense. According to the model, the central information to be assigned to a Semantic Unit is the **semantic type**. Each semantic type corresponds to a node in the SIMPLE Core Ontology, defined for all the languages, and it further specifies the information to be encoded for each entry belonging to this type, i.e. the semantic type is actually "a cluster of structured information".

In order to capture the multidimensional aspect of meaning, semantic information is organised in the form of Qualia Structure (Pustejovsky, 1995). Four roles are defined in Qualia Structure:

- Formal role places the entry in the appropriate node of a hierarchy
- **Constitutive role** provides information on the constitution of an entry
- **Telic role** provides information on the typical function(s) of an entity
- Agentive role concerns the origin of an entity, or its coming into being.
- The Qualia roles in SIMPLE are implemented as:
- *relations* holding between semantic units or *valued features*.

Each of the four Qualia roles is represented as a relation, which is in turn the top of a hierarchy of other more specific relations, representing more fine-grained subtypes of information of a given Quale. These hierarchies of relations within the four Qualia are referred to as *extended Qualia set*.

Each **relation** holds between the entry coded and one target Semantic Unit, where the schema [*Relation* (S1, S2)] signifies that there holds a relation between Semantic Unit S1 and Semantic Unit S2; in this schema, S1 is always the entry being coded, and S2 is the target Semantic Unit which must be specified. If for a certain relation there are more than one candidate target Semantic Units, the relation is repeated as many times as necessary, connecting S1 to a different S2 each time.

Features encode specific attributes of an entry; for each feature the model supplies a set of valid values – binary values or values specific to a feature (e.g. the feature *habitat* has three values, *earth*, *air* and *water*).

The model specifications include a set of relations and features, common to all languages. Moreover, each semantic type includes a set of mandatory and recommended (optional) relations and features to be coded, corresponding to the semantic information that is considered relevant to the particular type. This common principle adds to the uniformity of the content of semantic information included in the lexica across languages as well as to the uniformity of information coded for entries belonging to the same semantic type within one lexicon. For illustration purposes, Figure 2 shows the information to be encoded for the semantic type [Furniture].

Semantic Type:	[Furniture]
Formal:	<i>isa</i> (S1, <furniture> or <hypernym of="" s1="">)</hypernym></furniture>
Agentive:	created_by (S1, S2: [Creation])
Constitutive:	made_of (S1, S2) //optional//
	has_as_part (S1, S2) //optional//
Telic:	used_for (S1, <furnish>: [Event])</furnish>
	used_for (S1, S2: [Event]) //optional//

Figure 2: Example of the information to be coded for entries of the semantic type [Furniture]

4.2. Appropriateness of the CL to the application requirements

The four Qualia roles have a direct mapping to the constituents of the definitions that we have identified during our corpus analysis (Section 3.2). More specifically:

- the Formal Role, in the form of the *isa* relation, encodes the information <u>classifying</u> the entity in a hierarchy
- the Constitutive Role includes, in the form of relations (e.g. *made_of, has_as_colour, has_as_part,* etc.) or features (*dimension, habitat, state,* etc.), information on the properties of an entity
- the Telic Role provides information on the intended <u>functions</u> of an entity (e.g. object_of_the_activity, used_for, used_as, is_the_habit_of, etc.)
- the Agentive Role encodes information on the <u>provenance</u> of the entity (e.g. *created_by*, *derived_from*, etc.).

The systematicity of information contained in the desired definitions is catered for by the CL specifications, which dictate the types of information to be included in each encoded entry.

The structuring of the information in the desired definitions is performed by the Generator.

5. Description of the Generator

The Generator follows the pipeline architecture suggested by Reiter (1999), adapting, however, certain tasks to the application requirements. It consists of three main components:

- the Text Planner
- the Sentence Planner and

• the Surface Text Realiser.

The architecture of the system is shown in Figure 3.

The Text Planner is the module responsible for selecting the appropriate discourse plan, i.e. it handles the tasks of <u>content determination and organisation</u>, according to the discourse goal. In our application, the Text Planner determines the information elements that constitute the building blocks of a complete definition of concrete entities and retrieves from the knowledge base (i.e. the SIMPLE lexicon) relevant data required for each kind of information element.

Given the discourse goal of our application (medium-size dictionary definitions of concrete entities for native speakers), only one discourse plan is specified – referred to as "definition plan" hereafter. The Text Planner produces an XML list of Discourse Representations, which represent the content and the structure of the intended definitions in a language-independent formalism. A minimal set of rhetorical relations (Mann & Thompson, 1988) is used to join the different information elements into larger chunks. The mandatory information element for each and every object is the *identification message* mapped to the *isa* relation of the SIMPLE model, while all other information elements provide further details about the entity being defined, concerning its properties, functions and provenance.

The information elements required in order to formulate the definition of a concrete entity are semantically articulated in the four Qualia roles of our knowledge base, as presented in Section 4.2; therefore, the task of mapping the Discourse Representations to the SIMPLE Semantic Representations is rendered easier.

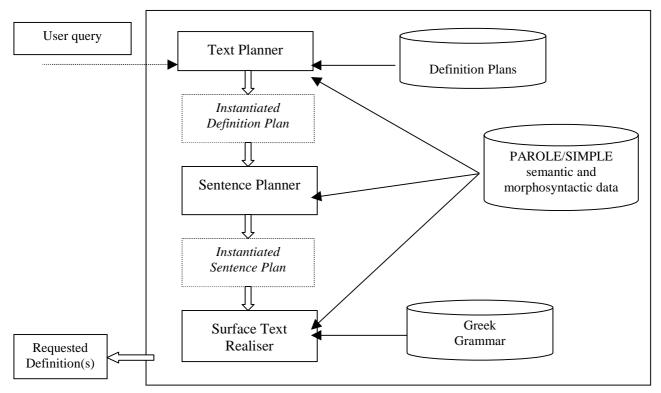


Figure 3: The system architecture

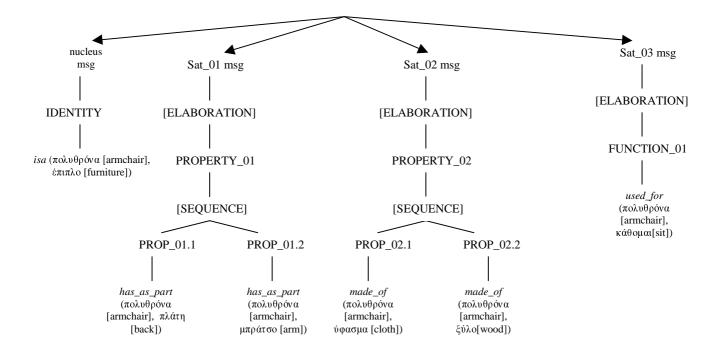


Figure 4: Example of an instantiated definition plan

The output of the Text Planner – an example of which is shown in Figure 4 – is realised as a tree structure populated by information elements at its leaf nodes¹.

The Sentence Planner converts the instantiated definition plan into a concise single-sentence definition, performing the tasks of <u>lexicalisation and aggregation</u> of the conveyed information.

The Sentence Planner maps the Semantic Representations (i.e. semantic relations, features and target Semantic Units) to the appropriate language specific lemmas and syntactic structures (e.g. for Greek, it generates the has_as_part relation as a PP governed by the preposition " $\mu\epsilon$ " [with], it renders the valued feature StateLiquid as an AP governed by the adjective "υγρός" [liquid], etc.). Moreover, by exploiting the link of the SIMPLE lexicon to the PAROLE morphosyntactic level, it extracts information that will be used at the surface realisation stage.

Aggregation caters for combining information elements into phrasal structures, using embedding and paratactic devices as appropriate: for instance, when the information elements are linked by a SEQUENCE rhetorical relation, conjoined phrases are used.

The Surface Text Realiser puts into a grammatically and orthographically correct sequence of words the content of the definition, i.e. it is responsible for the following:

- choosing the correct inflectional form of content words
- inserting function words, such as conjunctions and articles, in order to generate a fluent natural language text

- ordering words within the sentence
- applying orthographic rules.

The Surface Text Realiser receives input from the following resources:

- the morphosyntactic layer of the PAROLE/SIMPLE lexicon, which aims at producing the appropriate inflectional form for each target Semantic Unit, based on the morphosyntactic and lemma information it carries from the previous stages of the generation process
- a grammar module, which is application specific, given that definitions are written in a typical dictionary sublanguage
- a graphical component, which adds punctuation and converts the output into HTML format.

Figure 5 shows the final output of the Generator.

πολυθρόνα: έπιπλο με πλάτη και μπράτσα, που κατασκευάζεται από ύφασμα και ξύλο και χρησιμοποιείται για να καθόμαστε

[armchair: furniture with back and arms, which is made of cloth and wood and is used to sit on]

Figure 5: Sample output of the Generator

6. System evaluation

The advantages of using the above mentioned system architecture and linguistic resources for automatically generating dictionary definitions are the following:

• generation of **consistent and standardised definitions**: among the main problems cited in

¹ For illustration purposes, leaf nodes at the tree of Figure 4 contain, instead of the identification numbers, the lexicalised forms of the Semantic Units, which are normally inserted at the Sentence Planning stage of the generation.

literature concerning lexicography are the cyclicity of definitions and the use of unknown words; the Generator eliminates these problems since the output text is articulated in fluent but controlled natural language

- exploitation of a **well structured knowledge base**: the structure of the SIMPLE CL has proved itself well-tailored for our purposes, given that it contains information which is well defined, consistently encoded and organised in a hierarchical, standardised and rigorous way, thus facilitating the task of information extraction – for a discussion on these issues, see, for example, (Dale et al., 1998; Bateman, 1998)
- extensibility and portability of the system to other applications: given the architecture we have chosen for our system, it is possible to extend it without considerable changes in order to cover other similar applications; the enrichment of the system with additional user models and discourse goals (e.g. definitions for children's dictionaries) requires the addition of new discourse plans, rather than the re-design and construction of a new system from scratch
- exploitation of the CL as a multi-level lexical resource: the knowledge base of the application has direct access to the lexical resources (containing syntactic and morphological information) which are required for the Surface Text Realisation stage of the generation procedure, given that it forms part of a multilevel lexicon; thus, for the Surface Text Realisation task, no production of application specific lexical resources is required, and uniformity, consistency and full linguistic coverage (in terms of lexical items and their respective morphosyntactic features) is guaranteed
- catering for **multilinguality**: one of the advantages of using NLG techniques is the potential to generate texts in multiple languages; our system further enhances this advantage in two ways:
 - care has been taken in the architecture of the system so that specific language requirements are met at the later stages of the generation process
 - the resources we use (knowledge base and lexical resources) exist already for 12 languages, thus making it possible to generate definitions in all these languages with minimal changes in the system, since the model of the input remains the same; moreover, the CL model used can be extended to cover other languages as well.

Despite the undisputed advantages of the CL we are using for information extraction, two significant limitations have been identified:

• there is a **lack of information** in the CL, in comparison to the dictionary definitions we have studied in our corpus analysis: e.g. size information, or the number of parts constituting

a single entity is not given; this information can be crucial to the identification of a word sense

• there is **no indication as to the type of coordination that holds between relations of the same type**: for instance, a semantic unit can be made of (X AND Y), or of (X OR Y); this information is not formally represented in the SIMPLE model.

For the first flaw, a possible remedy is to add this information in the input. Given the structure of the SIMPLE lexicon, the addition of such information in the form of Qualia Structure, represented as features and/or relations, is possible without creating any problems.

As regards the second problem, the way we have already addressed it, is to add this information by parsing the input through a special interface that adds this information through an independent module, before the actual generation process.

7. Future work

Future work on the system is envisaged along the following lines:

- extending to definitions of other semantic types, which means that we will be faced with the extraction of more complex information present in the CL
- experimenting with other user models (e.g. children's dictionaries)
- experimenting with different definition plans: the possibility, for instance, to generate comparisons between entities that share similar characteristics and using this information to better illustrate the meaning of a word will be explored
- generating definitions in other languages, using the SIMPLE/PAROLE resources of these languages
- implementing the system as a hypertext application and making it available through Internet.

8. Bibliographical References

- Babiniotis, G. (1998). Λεξικό της Νέας Ελληνικής Γλώσσας [Dictionary of Modern Greek Language]. Athens: Kentro Lexicologias.
- Bateman, J. (1998). Deep Generation. In Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V. (Eds.), Survey of the State of the Art in Human Language Technology (pp. 151--155). Cambridge University Press.
- Boguraev, B. & Briscoe, E.J. (Eds.) (1989). Computational lexicography for natural language processing. London: Longman.
- Calzolari, N. & Briscoe, T. (1992). Acquisition of Lexical Knowledge from Machine-Readable Dictionaries and Text Corpora. In Proceedings of Coling-92, International Project Day, Giardini, Pisa.
- Calzolari, N. & Picchi, E. (1998). Acquisition of semantic information from an on-line dictionary. In D. Vargha (Ed.), Proceedings of COLING-88.
- Carenini, G., Pianesi, F., Ponzi, M. & Stock, O. (1993). Natural language generation and hypertext access. Applied Artificial Intelligence, 7, 135--164.

- Copestake, A. (1990). An approach to building the hierarchical element of a lexical knowledge base from a machine-readable dictionary. In Proceedings of the Workshop on Inheritance in Natural Language Processing (pp. 19--29).
- Cox, R., O'Donnell, M. & Oberlander, J. (1999). Dynamic versus static hypermedia in museum education: an evaluation of ILEX, the intelligent labelling explorer. In Proceedings of the Artificial Intelligence in Education conference.
- Dale, R., Green, S.J., Milosavljevic, M., Paris, C., Verspoor, C. & Williams, S. (1998). The Realities of Generating Natural Language from Databases. In Proceedings of the 11th Australian Joint Conference on Artificial Intelligence.
- Gavrilidou M., P. Labropoulou, E. Mantzari, S. Roussou, D. Anagnostopoulou & E. Desipri. (2000) Greek Lexicon Documentation, Simple LE4-8346, WP3.8.
- Gavrilidou, M., Labropoulou, P., Mantzari, E. & S. Roussou. (1998) Greek Lexicon Documentation, Parole LE2-4017/10369, WP3.9-WP-ATH-1.
- Instituto Neoellinikon Spoudon (1998). Λεξικό της Κοινής Νεοελληνικής [Dictionary of Koini Modern Greek]. Thessaloniki: Aristotle University of Thessaloniki.
- Kriaras, Emm. (1995). Λεξικό της Σύγχρονης Ελληνικής Δημοτικής Γλώσσας [Dictionary of the Contemporary Greek Dimotiki]. Athens: Ekdotiki Athinon.
- Mann, W.C. & Thompson, S.A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3), 243--281.
- Mellish, C., O'Donnell, M., Oberlander, J. & Knott, A. (1998). An architecture for opportunistic text generation. In Proceedings of the Ninth International Workshop on Natural Language Generation, Niagara-on-the-Lake, Ontario, Canada.
- Milosavljevic, M. & Dale, R. (1996). Strategies for Comparison in Encyclopaedia Descriptions. In Proceedings of the Eighth International Natural Language Generation Workshop (pp. 161--170).
- Neff, M. & McCord, M. (1990). Acquiring lexical data from machine-readable dictionary resources for machine translation. In Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in MT (pp. 87--92).
- Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, MA: The MIT Press.
- Reiter, E. (1999). Are Reference Architectures Integration Tools or Descriptive Aids? In Proceedings of the AISB-99 Workshop on Reference Architecture & Data Standards for NLP.
- Sanfilippo, A. & Poznanski, V. (1992). The acquisition of lexical knowledge from combined machinereadable dictionary sources. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy (pp. 80--87).
- SIMPLE Specifications Group (2000). Linguistic Specifications. SIMPLE LE4-8346, WP2, Deliverable D2.1.