# Narrative: Text Generation Model from Data

Darío Arenas[1], Javier Pérez[1], David Martínez[1],David Llorente[1], Eugenio Fernández[1], Antonio Moratilla[1],
[1](*Computer Science/ University of Alcalá, Spain)*

***ABSTRACT :*** *The generation of digital content has undergone a great increase in recent years due to the development of new technologies that allow the creation of content quickly and easily. A further step in this evolution is the generation of contents by automatic systems without human intervention. Thus, for decadesit has been developing models for the Natural Language Generation (NLG) that allow the transformation of content to the form of narratives. At present, there are several systems that enable the generation in text format. In this paper we present the Narrative system, which allows the generation of text narratives from different sources, and which are indistinguishable for user from those made by a human being.*

***KEYWORDS –****Automatic Digital Content Generation, D2T, NLG*

## I. INTRODUCTION

Nowadays, and in a general way, we understand digital content as every data, information, knowledge or a set of those, capable of being treated or stored by digital devices, and transmitted through telecommunication networks. The format of that digital content is usually categorized basically in four types or formats: text, image, audio and video; and if web pages, blogs social media, or digital newspaper could be considered in technical terms as sets of elements from those four basics elements, eventually they start to treat it as per se formats; for other digital content as software in general or videogames where the concept is not that evident define by those basics formats. Regardless of this kind of classification, the creation of digital content of every type has suffered an exponential increase in last decades.

This tremendous growth in digital content generation has been affected positively by different factors. On the one hand, we have the fast evolution in hardware and software which has allowed the creation of content faster, easier and more efficient each time. On the other hand, digital content has arrived to everyone at final user level, this implies that each day more people are using technology and it has been developing more complex content to satisfy that public, such as web generation, and treatment of images, audios or videos. Besides, there are social factors which have allowed this huge increase in digital content generation, as we were saying, like the contents demanding from users which have even created many moments of saturation in certain content.

From now on we will be referring to text digital content as information, because it is the approach of this work. That kind of information which digital media publish for users consuming in digital newspaper, blogs or webs for electronic commerce. For example, those pieces of information can be a journalist article, a product description, a brief event explanation, a set of rules or the explanation of weather predictions. In this way, in this work, the subset of digital content we are going to deal with is digital information in text format. However, the model of automatic generation we are going to expose in this work is perfectly valid for other different subsets of digital content, not only for texts ones.

## II. NATURAL LANGUAGE PROCESSING

Automatic generation of information in text format is usually approached from Natural Language Processing (NLP in what follows), which combines techniques from Artificial Intelligence, like Logic or Machine Learning, with Statistics, Applicated Linguistics or Procedural Programming, in order to develop methods that allow us tasks such as understanding and computer assisted processing of information given in human languages for certain jobs, like automatic translation, narratives generation, knowledge extraction from texts, etcetera. NLP is divided in

two large areas, Natural Language Generation (NLG hereafter) and Natural Language Understanding (NLU hereafter), which can be considered as the inverse task of NLG. On the one hand, NLG aim is to generate text messages in human language from structure and complete data of the domain we want to describe. On the other hand, NLU aims is obtain structure and understanding data, given texts in human language. In this work, we are keen on generate narratives automatically, so we are interested in NLG. It is not a new discipline, and it has support from other disciplines which implies a lot of techniques from those fields. NLP was born in the 1960's, although it was not recognize until the decade of 1980 [1],[2],[3],[4].

From the works of different researchers, principally Reiter's and Dale's ones (1997), it has been developing a model in recent years for solving NLG problems, it is based on sequentially and well-define stages ([5],[6],[7],[8],[9], [10],[11],[12],[13],[14],[15],[16],[17],[18]), from which it has developed specific models to apply to a large number of fields, such as meteorology, health or industrial processes [19].

This model is based on six phases or activities grouped in three stages, so each one works in different levels of linguistic representation (semantic, lexicon, syntactic): Text Planning (activities 1 and 2); Sentences Planning (activities 3, 4 and 5); Linguistic Realization (activity 6). Those six basics activities which are represented on Fig. 1 are [20]:

- Contents determination: in this first activity, from given data, we generate a set of simple messages summed up in an intermediate language which distinguishes labels, objects, concepts and relations of interest in the domain of application, and they stablish the information that will be on the text.
- Speech Planning: we order and structure the messages to transmit.
- SentencesAggregation: we group the messages on sentences to improve text fluency.
- Lexicalization: we choose the specific words and expressions that we will use to express the concepts and relations on the domain.
- Generation of reference expressions: we select word or expression which identify objects from the domain, in that way, we can ensure that the system provides enough information to distinguish each object from the others.
- Linguistic Realization: we apply grammatic rules to provide a right final text fromlexicon, syntactic and semantic point of view.

present objective information with no emotional

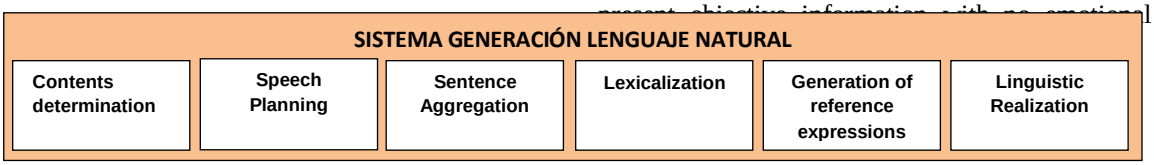


**Figure 1**. Components of a NLG system.

## III. TEXT GENERATION FROM DATA

At present, the different developed systems based on that model generate two kinds of texts. On the one hand, those we can name "objective texts", that present information about a well-defined domain (sports, finance, meteorology, products description, etcetera) and those that usually define the domain to deal with, in terms of data. On the other hand, there are "subjective texts", in which we express stories with a high semantic weight and they have a lot of emotionalconnotations, in contrast with objective text which rarely get it because they

In this work we will focus on objective text, inwhich we find a well-defined domain from a finiteset of data. In the last years, that has been named D2T models (data-to-text models). It has increased the interest of NLG researchers [21] due to the explosion in data generation, and the easy access to it and their treatment. Therefore, researchers have appreciated a practical application for those techniques that have been developing for the last years, in order to try to create systems which permit to generate natural language

successfully, that have been applied with more or less success to task like automatic translation of languages. Our goal is to achieve an approximation to the model in Fig. 1 in a context where data are the content definition, several D2T models in the last years have been developed, and now, there is a consensual structure followed by the vast majority of practical implementation [22].

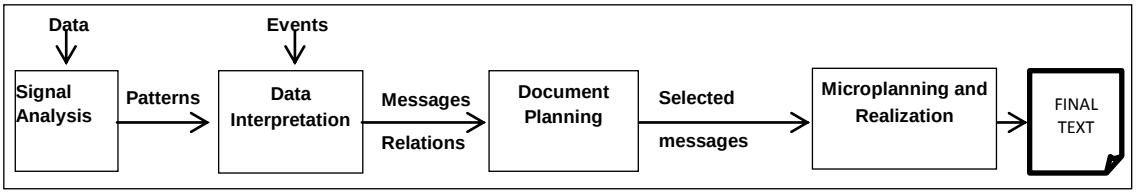


**Figure 2**. Components of "data-to-text" NLG component system.

This model (Fig. 2) is based on 4 activities: Signals Analysis (it analyzes given data looking for patterns and trends); Data Interpretation (it identifies complex messages in the domain from patterns and trends found in the first activity, identifying relations among the messages); Document Planning (it decides which messages must appear in the final text); Microplanning (it transforms messages into several linked expressions) and Realization (those expressions are transformed into text strings). Later, it has been modified and used by different authors generatingsimilar approximations, like in [18] which propose those activities: Data Analysis, Content Defining (select and structure the information given from the data analysis that we want to transmit), Aggregation and Microplanning (it transforms information into linked expressions), and Realization (those expressions are transformed into text strings). With that model structure as base,specific D2T models have been developed like WeatherReporter, FoG[5], MultiMeteo[23], SumTime-Mousam [24], British Met Office [25], TEMSIS [26], and MARQUIS [27] in meteorology field; SUREGEN-2 [28] and BabyTalk[21] in health field, Patent Claim Expert [29] y Sum-Time-Turbine [30] in industry field; IDAS [31] to automatic generation of online documents; ModelExplainer[32] to generate text description of software models oriented to objects; PEBA [33] to entity description in a knowledge base; STOP [34] to automatic generation of personal letters to give up smoking. In relation with generation systems for subjective text there are several systems like TALE-SPIN [35],GESTER [36], MINSTREL [37], MEXICA [38], BRUTUS [39], Cast [40].

If we analyze those and other implemented systems, beyond minimal difference with the model in Fig. 2 to obtain a specific D2T model, we can classify those system in two types, those based on writing scripts, and those which do not use it. From a formal perspective we can say that both system should get similar solutions in quality terms, it is truly that there are some practical aspects that makeit impossible to obtain successful results without using writing scripts, such as execution time, aspects related with journalistic style, semantic aspects like ambiguities, and those related with emotional charge in the messages. Given those aspects is rather difficult to get automatic messages with D2T systems. Anyway, authors as Van Deemter[41] have analyzed and compared those two types of systems and have concluded that there are no difference in quality results in system based on writing scripts.

Another relevant aspect in development of that kind of system, is the validation of obtaining results and we do that by using evaluation methods to check if they accomplish the requirements a priori, which is not easy, because there are requirements difficult to express and quantify, like those related with the style in narratives. Others as users or consumer acceptation can not be established until checking if the obtained results satisfy their necessities. Eventually, it is subject of study and that is a consensual thought in the community of researchers. In that way, we found that the vast majority of known methods for evaluation are quantitative ([42],[43]) and they try to obtain a numeric measure about quality (making questionnaires to human experts; similarity measures among generated texts and representative texts, etcetera). There are qualitative methods used to identify and to correct some aspects in content analysis stages and in speech stages ([44],[45]).

## IV. NARRATIVE MODEL

From an applied perspective, the development of D2T system is justified by the constant increase on contents demanded by users (information consumers), eventually this implies situations in which it is impossible to generate that amount of information with enough quality because we do not have the human resources to make it possible. Thus, it is necessary to ask ourselves in a general way, if it is possible to automatize text generation with developed techniques and with enough quality, that is, showing complete and accurate information, grammatically correct, with a proper style, etcetera, which satisfy consumers requirements about the content.

To achieve an illustration of our problem, we are going to describe three real examples solved by Narrative (www.narrativa.com) as an answer to users demanding, in that case companies, which are looking for solutions to their digital content generation problems in text format:

- First one is about sport information offered by written media. In all terms, a digital written media can offer high quality information to users about more relevant matches in principal leagues of popular sports. However, when it is a minority sport or they are lower leagues, offering high quality information to users is non-viable in economics terms (understanding high quality information as complete an accurate information), because that requires some task by the professionals as covering all matches that is economically non-viable.
- Second example comes from the necessity of giving a quality description about a huge quantity of products that offer some webs for online shopping. In this case, quality concept acquires more importance because we do not achieve it only by enumerating or simply describing characteristics of the products, we need appropriate descriptions for each product in a specific way departing from their characteristics and other data, so that this generated information could identify completely and accurately every product among others similar, providing useful information to users or consumers when taking right decisions. This kind of information is usually described by an expert on the area, but for millions of products, it is non-viable for human experts to generate that much information on time and shape.
- The third one deals with the problematic of these situations that need a quick text generation and it is impossible for humans to generate on time. For instance, task execution orders for people, which are in base of data, and must be modified in seconds, such as products delivery planning. Simply, we have to imagine a group of truck drivers loading product into the truck. In headquarters someone is planning the load and the delivery of the products, from that moment, drivers have to receive text orders in voice format, which indicate them at real time the route they must be keeping (the route will be modified depending on traffic, business interest, etcetera…) and some other useful information.

Those examples, which can be extrapolated it to similar situations, illustrate a current situation in which users demand information with different quality degrees, and that is impossible to generate on time and shape, because it requires human participation in the generation process.It has an effect on business models, which make them non-viable and generate the lack of certain useful information demanded by users at presents. In that situation, what matters, if possible, is automatize, that is, to develop automatic systems that allow to generate that information with enough quality to be offered to users and consumers. Narrative's system answers that question showing the possibility to generate information in text format like narratives, from sets of representative data from a specific domain, such as some of those commented before.

To design our model, we start from the issue that generating text in natural language from given data, we must develop systems which have to face three phases (Dara Analysis, Content Determination and Realization) with high complexity in scientific terms, but eventually is possible to develop approximate methods, principally from Big Data techniques, statistics and Artificial Intelligence, which permit to obtain useful solutions and acceptable to practical endings.

First phase is related to data treatment, that is, determine, obtain, analyze, classify and categorize necessarily data needed to obtain the expected narratives with the required quality, that is, complete and accurate narratives [3]. Data must be expressed in data bases and in knowledge bases which will lead to proper treatments later.

As a general rule, the vast majority of development system until now approach given data to the model in plain text, that is, they work exclusively with structure data, generally numeric. However, as we can see it in the model displayed and in Fig. 3, this concept to another and more complex given data (documents with simple data, documents elaborated in grammatical and emotional aspects, images, audios, videos, social media, emails, web pages, etcetera) from an analysis and extraction data module, which obtain known and relevant information with Big Data techniques, with the goal of getting a final domain represented in a specific knowledge base, undergoing through an intermediate process of data bases generation that will allow to aggregate proper data of the client, to realize an unification process, and to keep a format to extrapolate to any other model of knowledge base later.

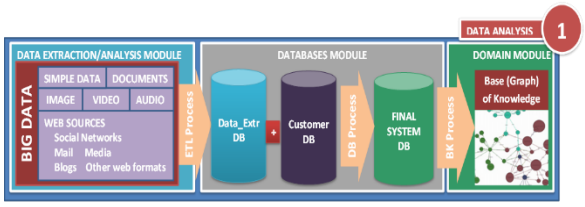


**Figure 3**. Components of Data Analysis phase.

In second phase (Fig. 4), from the defined domain in previous phase, and including client requirements and other external agents which can conditionate the final information, then it extracts the relevant information that is subjected to appear in the final text, using new analysis and exploring techniques, principally of Artificial Intelligence, about the knowledge base in that case.


**Figure 4**. Components of Content Determination phase.

Finally, once it has been obtained the relevant information to express in the final text, it realizesone and last phase to conform the final text so that it fulfils all the editorial requirements from the media or the client, such as style or proper aspects of the language (Fig. 5).


**Figure 5**. Components of Realization phase.

In parallel with second and third phases, the model is based on the use of written scripts at different levels of granularity. It is known that humans when generating text information are capable of managing with a huge amount of information and knowledge related with common sense, modelling information and preparing messages adding subjective aspects related with context or mood. Thus, automatic text generation presents lack of naturality an emotion, and poverty in semantic content, when comparing with human texts. Scientifics propose that, nowadays, humans are capable of making complete and more complex content than generating by automatic systems based on NLG techniques, and it will last years, probably decades, until obtaining automatic approximations similar in quality terms.

Nevertheless, we hold that, from an applied perspective, there is a way to deal with that inconvenient. In those applied fields, there are experts capable of generating narratives with those complex aspects mentioned, and that the automatic system is unable to show them in the final text. It is possible to realize a process of atomic written scripts, representatives of the domain, structures at different levels of granularity, which add all the emotional, subjective and style aspects, which an automatic system is unable of. The module of atomic written scripts is used as base to train the system and to adapt to final texts, so that they show information similar to a text generated by experts.

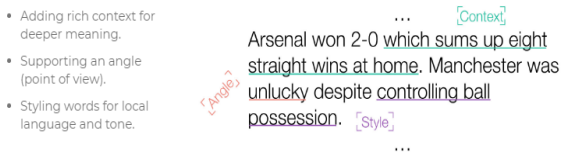


**Figure 6**. Narrative's text example from given data.

Narrative's model is based on those three stages shown in Fig. 1, 2 and 3 and on the module for managing the written scripts. It is shown in Fig. 6 a descriptive text of a football match summary generated automatically by the system from Premier League provided by the Company Opta (www.optasports.es).

## V. CONCLUSION.

The development of automatic systems for narratives generation in text format has undergone a constant development in the last decades, in research of all the aspects related to natural language. Nevertheless, we are far from showing subjective aspects like emotions and complex semantic of human texts in automatic generation systems. However, eventually, it is possible by using the advantages provided in that field and tools like written scripts, to build automatic useful systems which allow massive generation of contents on real time, which is impossible only with human intervention. In addition, they have characteristics that make them indistinguishable from human ones. In this work, we present the Narrative's system which correspondsto that philosophy, and which is used at present to solve practical problems in different real situations of content production for objective public who consume it through web platforms.

## REFERENCES

[1] *Goldman, N. (1975). Conceptual generation. En R.C. Schank, Conceptual.*

[2] *Davey, A. C. (1978). Discourse production. Edinburgh: Edinburgh UP.*

[3] *McKeown, K. R. (1985). Discourse strategies for generating natural-language text.*

[4] *Appelt, D. E. (1985). Planning english sentences. Cambridge University Press.*

[5] *Goldberg, E., Driedgar, N., & Kittredge, R. (1994). Using natural-language processing to produce weather forecasts. IEEE Expert, 9, 45-53.*

[6] *Dalianis, H. (1996). Aggregation as a subtask of text and sentence planning. Proceedings of the 9th Florida Artificial Intelligence Research Symposium (FLAIRS-96), 1-5. Florida (EEUU), 20-22.*

[7] *Not, E. (1996). A computational model for generating referring expressions in a multilingual application domain. Proceedings of the 16th International Conference onComputational Linguistics (COLING'96). 2, 848-853. Copenague (Denmark).*

[8] *Reiter, E. & Dale, R. (1997). Building Applied Natural-Language Generation Systems. Journal of Natural-Language Engineering. 3, 57-87.*

[9] *Bateman, J.A. (1997). Enabling technology for multilingual natural language generation: the KPML development environment. Journal of Natural Language Engineering, 3, 15-55.*

[10] *Cheng, H. & Mellish, C. (1997). Aggregation based on text structure for descriptive text generation. Proceedings of the PhD Workshop on Natural Language Generation, 9th European Summer School in Logic, Language and Information (ESSLLI97). Aix-en-Provence (France), 18-22.*

[11] *Shaw, J. (1998). Clause aggregation using linguistic knowledge. Proceedings of the 9th International Natural Language Generation Workshop (INLG'98). Canada. 138-147.*

[12] *Cahill, L. &Reape, M. (1999). Component tasks in applied NLG systems. Technical Report ITRI-99-05. Information Technology Research Institute, University of Brighton (United Kingdom).*

[13] *Teich, E. (1999). Systemic functional grammar in natural language generation: Linguistic description and computational representation. Cassell Academic Publishers,Londres (United Kingdom).*

[14] *Theune, M. (2000). From data to speech: language generation in context. Tesis doctoral, Eindhoven University of Technology (PaísesBajos), 2000*

[15] *Reiter, E. & Dale, R. (2000). Building Natural Language Generation Systems. Cambridge University Press, Cambridge.*

[16] *Bangalore, S. &Rambow, O. (2000). Corpus-based lexical choice in natural language generation. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000), (pp. 464- 471). Hong-Kong (China).*

[17] *Díaz-Hermida, F, Ramos-Soto, A. &Bugarín, A. (2011). On the role of fuzzy quantified statements in linguistic summarization. En Proceedings of 11th International Conference on. Intelligent Systems Design and Applications (ISDA), 166–171.*

[18] *Hunter, J., Freer, Y, Gatt, A., Reiter, E., Sripada, S. & Sykes, C. (2012). Automatic Generation of Natural Language Nursing Shift Summaries in Neonatal Intensive Care: BT-Nurse. Artificial intelligence in medicine. 56(3), 157-172.*

[19] *Bateman, J.A. (2001). Natural language generation: an introduction and open-ended review of the state of the art, http://www.fb10.unibremen.de/.*

[20] *Bautista, S. (2008). Generación de textos adaptativa a partir de una elección léxica basada en emociones. Trabajo de Máster en Investigación en Informática. Universidad Complutense de Madrid. http//eprints.ucm.es/10061/1/ProyectoFinMaster.pdf.*

[21] *Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y. & Sykes, C. (2009). Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. Artificial Intelligence. 173(7). 789-816.*

[22] *Reiter, E. (2007). An architecture for data-to-text systems. Proceedings of the 11th European Workshop on Natural Language Generation (ENLG). Germany. 97-104.*

[23] *Coch, J., Dycker, E.D., García-Moya, J.A., Gmoser, H., Stranart, J.F., &Tardieu, J. (1999). Multimeteo: adaptable software for interactive production of multilingual weather forecasts. En Proceedings of the 4th European Conference on Applications ofMeteorology (ECAM 99), Norrk¨oping, Sweden).*

[24] *Sripada, S., Reiter, E., & Davy, I. (2003): Sumtimemousam: Configurable marine weather forecast generator. Expert Update. 6(3), 4–10.*

[25] *Sripada, S.G., Burnett, N., Turner, R., Mastin, J., & Evans, D. (2014): A case study: Nlgmeeting weather industry demand for quality and quantity of textual weather forecasts. En INLG-2014 Proceedings.*

[26] *Busemann, S. &Horacek, H. (1997). Generating air-quality reports from environmental data. EnBusemann, S., Becker, T., Finkler, W., eds.: DFKI Workshop on Natural Language Generation, DFKI Document D-97-06.*

[27] *Wanner, L., Bohnet, B., Bouayad-Agha, N., Lareau, F., &Nicklass, D. (2010). Marquis: Generation of user-tailored multilingual air quality bulletins. Applied Artificial Intelligence. 24(10), 914–952.*

[28] *Hüske-Kraus, D. (2003): Suregen 2: A shell system for the generation of clinical documents. En Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003). 215–218*

[29] *Sheremetyeva, S., Nirenburg, S., &Nirenburg, I. (1996): Generating patent claims from interactive input. En Proceedings of the 8th. International Workshop on Natural Language Generation (INLG'96), Herstmonceux, England. 61–70.*

[30] *Yu, J., Hunter, J., Reiter, E., &Sripada, S. (2001) An approach to generating summaries of time series data in the gas turbine domain. En Proceedings of IEEE International Conference on Info-tech & Info-net (ICII2001). Beijing. 44–51.*

[31] *Reiter, E., Mellish, C., & Levine, L. (1995). Automatic generation of technical documentation. Applied Artificial Intelligence. 9, 259-287.*

[32] *Lavoie, B. & Owen, R. (1997). A fast and portable realizer for text generation. Proc. of the Fifth Conference on Applied Natural-Language Processing. 265-268.*

[33] *Milosavljevic, M. & Dale, R. (1996). Strategies for comparison in encyclopedia descriptions. Proceedings of the Eighth International Workshop on Natural-Language Generation. 161-170.*

[34] *Reiter, E., Robertson, R., & Osman, L. (1999). Knowledge acquisition for natural language generation. En Proceedings of the Joint European Conference on Articial Intelligence in Medicine and Medical Decision Making. 389-399.*

[35] *Schank, R.C. (1969). A conceptual dependency representation for a computer-oriented semantics. Phd thesis, University of Texas.*

[36] *Pemberton, L. (1989). A modular approach to story generation. En 4th European Conference of the Association for Computational Linguistics. Manchester.*

[37] *Turner, S.R. (1992). Minstrel: A computer model of creativity and storytelling. Informe Técnico UCLA-AI-92-04, Computer Science Department, University of California.*

[38] *Pérez y Pérez, R. & Sharples, M. (2004). Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. Knowledge Based Systems Journal, 17(1), 15-29.*

[39] *Bringsjord, S &Ferrucci, D. (1999). Artificial Intelligence and Literary Creativity. En Inside the mind of Brutus, a StoryTelling Machine. Lawrence Erlbaum Associates, Hillsdale, NJ.*

[40] *León, C. &Gervás, P. (2008). Cast: Creative storytelling based on transformation of generation rules. En P. Gervás, R. Pérez y Pérez, y T. Veale, editores, Proceedings of the 5th International Joint Workshop on Computational Creativity.*

[41] *Van Deemter, K., Krahmer, E. &Theune, M. (2005). Real Versus Template-based Natural Language Generation: a False Opposition. Computational Linguistics. 31(1), 15-24.*

[42] *Belz, A. & Reiter, E. (2006). Comparing automatic and human evaluation of nlg systems. In Proc. European Conference Association for Computational Linguistics (EACL'06). 313–320.*

[43] *Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). Bleu: A method for automatic evaluation of machine translation. En Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02. Stroudsburg, PA, USA, 311–318.*

[44] *Sambaraju, R., Reiter, E., Logie, R., McKinlay, A., McVittie, C., Gatt, A., &Sykes,C. (2011). What is in a text*

*and what does it do: Qualitative evaluations of an nlg system the bt-nurse using content analysis and discourse analysis. En Proceedings ofthe 13th European Workshop on Natural Language Generation. 22 – 31*

[45] *Reiter, E. (2011). Task-based evaluation of nlg systems: Control vs real-world context. En Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop. UCNLG+EVAL'11, Stroudsburg, PA, USA. 28–32*