INLG 2016

The 9th International Natural Language Generation conference

Proceedings of the Conference

September 5-8, 2016
Edinburgh, UK

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Introduction

Welcome to the 9th International Natural Language Generation Conference (INLG 2016)! INLG is the biennial meeting of the ACL Special Interest Group on Natural Language Generation (SIGGEN). The INLG conference provides the premier forum for the discussion, dissemination, and archiving of research and results in the field of Natural Language Generation (NLG). Previous INLG conferences have been held in Ireland, the USA, Australia, the UK, and Israel. Prior to 2000, INLG meetings were held as international workshops with a history stretching back to 1983. In 2016, INLG is organized in Edinburgh, UK.

The INLG 2016 program includes presentations of substantial, original, and previously unpublished results on all topics related to natural language generation. This year, INLG received 65 submissions (29 full papers, 30 short papers and 6 demo proposals) from 12 different countries. 16 submissions were accepted as full papers (10 presented orally, 6 as posters), 20 as short papers (4 presented orally, and 16 as posters), and 3 demos. In addition, INLG 2016 includes an invited talk by Yejin Choi (University of Washington) and Vera Demberg (Saarland University).

The organizing committee would like to offer their thanks to our invited speakers for agreeing to join us and to the authors of all submitted papers. We have also received sponsorship from Arria and WebNLG, for which we are extremely grateful. Finally, we would like to welcome you to Edinburgh and hope that you have an enjoyable and inspiring visit!

Amy Isard, Verena Rieser and Dimitra Gkatzia
INLG 2016 Co-Chairs

SimpleNLG-IT: adapting SimpleNLG to Italian

Alessandro Mazzei and **Cristina Battaglino** and **Cristina Bosco**
Dipartimento di Informatica
Università degli Studi di Torino
Corso Svizzera 185, 10149 Torino
[mazzei,battagli,bosco]@di.unito.it

Abstract

This paper describes the SimpleNLG-IT realiser, i.e. the main features of the porting of the SimpleNLG API system (Gatt and Reiter, 2009) to Italian. The paper gives some details about the grammar and the lexicon employed by the system and reports some results about a first evaluation based on a dependency treebank for Italian. A comparison is developed with the previous projects developed for this task for English and French, which is based on the morpho-syntactical differences and similarities between Italian and these languages.

1 Introduction

Natural Language Generation (NLG) involves a number of elementary tasks that can be addressed by using different approaches and architectures. A well defined standard architecture is the pipeline proposed by Reiter and Dale (Reiter and Dale, 2000).

In this approach, three steps transform raw data into natural language text, that are: document planning, sentence-planning and surface realization. Each one of these modules triggers the next one addressing a distinct issue as follows. In document planning the user decides the information content of the text to be generated (*what to say*). In sentence-planning, the focus is instead on the design of a number of features that are related to the information contents as well as to the specific language, as the choice of the words. Finally, in surface realisation, sentences are generated according to the decisions taken in the previous stages and by fulfilling the morpho-syntactic constraints related to the language specific features, like word order, inflection and selection of functional words.

Surface realisers can be classified on the basis of their input. Fully fledged realisers accept as input an unordered and uninflected proto-syntactic structure enriched with semantic and pragmatic features that are used to produce the most plausible output string. OpenCCG is a member of this category of realisers (White, 2006). Indeed, OpenCCG accepts as input a semantic graph representing a set of hybrid logic formulas. The hybrid logic elements are indeed the semantic specification of syntactic CCG structures defined in the grammar realiser. The semantic graph under-specifies morpho-syntactic information and delegates to the realiser many lexical and syntactic choices (e.g. function words). Chart based algorithms and statistical models are used to resolve the ambiguity arising from under-specification.

In contrast, *realisation engines* are simpler systems which perform just linearisation and morphological inflections of the proto-syntactic input. As a consequence, realization engine presumes a more detailed morpho-syntactic information as input. A member of this category is SimpleNLG (Gatt and Reiter, 2009). It assumes a complete syntactic specification, but unordered and uninflected, of the sentence in the form of a mixed constituency/dependency structure. Content and function words are chosen in input as well as modifiers order. The greatest advantage of this system is its simplicity, which allows to pay more efforts in the previous stages of the NLG pipeline.

SimpleNLG was originally designed for English but it has been successively adapted to German,

Proceedings of The 9th International Natural Language Generation conference, pages 184–192,
Edinburgh, UK, September 5-8 2016. ©2016 Association for Computational Linguistics

French, Brazilian-Portuguese and Telugu (Bollmann, 2011; Vaudry and Lapalme, 2013; de Oliveira and Sripada, 2014; Dokkara et al., 2015). The first contribution of this paper is the adaptation of SimpleNLG for Italian[1]. The most challenging issues under this respect of this project (see Sections 2 and 3) are: (1) the Italian verb conjugation system, that cannot be easily mapped to the English system and shows many idiosyncrasies; (2) the high complexity of the Italian morphological inflections; (3) the lack of a publicly available computational lexicon suitable for generation. Nevertheless, the contribution of this paper goes beyond the adaptation of the existing implementation to a novel language. We applied indeed a treebank-based methodology (see the monolingual and multilingual resources cited below) for both evaluating our results (see sec. 4), and describing in a comparative perspective the features of the implemented grammar, referring to the differences between Italian, French and English. This makes the work more linguistically sound and data-driven. We started our work from SimpleNLG-EnFr1.1, that is an adaptation to French (Vaudry and Lapalme, 2013) of the model developed for English in (Gatt and Reiter, 2009). A property of our project is multilingualism: by using the same architecture of SimpleNLG-EnFr1.1 we are able to multilingual documents with sentences in English, French and Italian.

In porting SimpleNLG-EnFr1.1 to SimpleNLG-IT, we created 10 new packages and modified 28 existing classes. The morphology and morphonology processors needed to be written from scratch because of the features that differentiate Italian with respect to French and English. The Syntax processor needed to be adapted, especially for the management of noun and verb phrases and for clauses. However, at this stage, we used the same orthography processor of French. We needed to extend the system with 33 new lexical features, necessary for accounting verb irregularities (subjunctive, conditional, remote past, etc.) and for processing the superlative irregular form of the adjectives.

In the next Sections we survey the main features of SimpleNLG-IT, in particular: in Section 2 we describe the grammar defined by the system, that has been developed starting from the SimpleNLG-EnFr1.1. grammar; in Section 3 we describe the lexicon adopted, that has been built starting from three lexical resources available for Italian; Section 4 describes the evaluation of the system, which is based on examples from both grammar books (Patota, 2006) and an Italian treebank (Nivre et al., 2016); finally, Section 5 closes the paper with some final considerations and pointing to future works.

2 From French to Italian grammar

In this Section we will focus on the generation of constituents and on their order within the sentence in Italian. In this achievement, the main reference for Italian grammar is (Patota, 2006). In general, it must be observed that Italian, like French, is featured by a rich inflection that is clearly attested by Verbs, but also by the behavior of other grammatical categories whose *morphosyntactic* features (e.g. gender, number and case) are crucial for determining their syntactical order in the phrases to be generated. As stated above, our approach is based on that adopted for French in (Vaudry and Lapalme, 2013), which has been in turn inspired by that used for English (Gatt and Reiter, 2009). First of all, we developed therefore a comparison among Italian and these other two languages in order to detect the main novel features to be taken into account in the development of SimpleNLG-IT. The parallel treebank ParTUT[2] developed for Italian/French/English helped us in this comparison.

In the rest of this Section we organize these features in the main classes which did drive the processes we implemented: morphology and syntax, which are strictly interrelated because of the concordance phenomena, and morphonology.

2.1 Morphology and syntax

2.1.1 Verb conjugation

Italian is featured by a complexity of inflection which is typical of morphologically rich languages and its richness, in this perspective, positively compares with that of French. Nevertheless, in order to

[1] SimpleNLG-IT: `https://github.com/alexmazzei/SimpleNLG-IT`

[2] `http://www.di.unito.it/~tutreeb/treebanks.html`

develop a suitable model for Italian verbs, we differentiate the implementation of SimpleNLG-IT under this respect with that exploited in SimpleNLG-FrEn1.1.

The main traits that we have assumed in this phase of the project for modeling verbs are tense, progressive and perfect, as can it be seen in the Table 1. The opposition between the different features, i.e. perfect and imperfect, can be expressed by using different means in different languages. While in English aspect is especially relevant and strictly interrelated with mood and tense, in Italian and other Romance languages derived from Latin several means are available for expressing it, which vary from inflection, to lexical selection, to syntactic choice of periphrastic forms and a system of moods richer than that of English. On the one hand, the imperfect forms for present (I'm writing) and past (I was writing) and the perfect forms for present (I have written) and past (I had written) exploited in English cannot always find a unique correspondence in Italian forms. On the other hand, while the progressive form *Io sto scrivendo* surely corresponds to *I'm writing*, the form *Io scrivo* can be translated with *I write* or *I'm writing*, and the second selection is preferred in particular when a modifier is associated with the verb, like in *Io scrivo in questo momento* (I'm writing in this moment).

In order to reproduce the complete Italian verb conjugation system, we used the features TENSE[3], PERFECT, PROGRESSIVE (Table 1). Moreover we used the feature FORM to set the tenses gerund, infinitive, subjunctive.

2.1.2 Noun phrase construction

The noun phrase may include, beyond the noun, also specifiers (i.e. determiners) and modifiers (i.e. adjectives and adverbs). For specifiers the main issue to be dealt with consists in setting their morphosyntactic features according to those of the noun, assuming that their position within the noun phrase is before the noun and the premodifiers. It can be observed that Italian is more similar to French than to English for what concerns specifiers, since in most of cases nouns are mandatorily associated with specifiers, while English nouns often occur without determiners.

Italian conjugation	Tense	PE	PR
indicativo presente	present	F	F
imperfetto	past	F	F
futuro semplice	future	F	F
futuro anteriore	future	T	F
passato prossimo	past	T	F
passato remoto	remote-past	T	F
trapassato prossimo	plus-past	T	F
trapassato remoto	plus-remote-past	T	F
passato remoto	remote-past	T	F
presente progressivo	present	F	T
passato progressivo	past	F	T
futuro progressivo	future	F	T
condizionale presente	present	F	F
condizionale passato	past	F	F
congiuntivo presente	present	F	F
congiuntivo imperfetto	past	F	F
congiuntivo passato	past	T	F
congiuntivo trapassato	plus-past	T	F

Table 1: Relation between verb tenses and traits in Italian: **TENSE** is a multi-value feature; **PE**rfect and **PR**ogressive are two boolean features.

The canonical NP word order is spec > preMod > noun > complements > postMod, but we need to introduce a number of new lexical features to account for the peculiar adjective word order with adjective types. The position that an adjective assumes with respect to the noun varies indeed accordingly with its type: ordinal, possessive and qualitative adjectives usually precede the associated noun, while colour, geografic and relation adjectives behave as noun's postmodifiers. See, e.g., *la grande casa gialla* (the big yellow house) where the adjective *big* is a qualitative adjective while *yellow* is a colour adjective. Moreover, when more than one adjective occurs, like a pre or postmodifier, a specific order must be respected, e.g. possessive > ordinal > qualitative is the canonical order for premodifiers. See e.g. *il mio primo grande viaggio* (my first big travel).

Finally, similar to SimpleNLG-EnFr1.1 we treated interrogative and demonstrative adjectives as specifiers, in contrast to the reference grammar book, which considers them as modifiers.

[3] We add the values simple_past, remote_past, plus_past, plus_remote_past as possible values of the feature TENSE.

2.1.3 Verb phrase and sentence construction

Among the main features to be taken into account in generating a sentence there is the order of constituents, which can also strongly vary according to language and typology of sentence. For what concerns Italian, the word order in declarative sentences, as reported in the study based on the parallel treebank ParTUT developed for Italian/French/English (Sanguinetti et al., 2013), is featured by a larger variability with respect to the other two languages, since the SVO order is detected in 74.5% of Italian sentences, in 82.4% of French sentences and in 88.5% of the English ones. Nevertheless, observing that SVO is usually tolerated in Italian in most of cases, at least for the purpose of practical NLG applications, the SVO order can be exploited. The conventional word order adopted by SimpleNLG-IT in the construction of the verbal phrase is `auxiliarie(s)` > `premod` > `verb` > `premod` > `complements` > `postmod` where the order of the complements is `direct-object` > `indirect-object` > `other-complements`. See e.g. *ho spesso dato libri a Mario in regalo* ([I] often gave books to Mario as present).

2.1.4 Negative sentences

In French, negative sentences are featured by the canonical presence of the adverb *pas* after the verb negated by the adverb *ne* (not). For instance in *Je ne mange pas les pommes* (I don't eat apples). In Italian the negation adverb *non* (not) precedes the verb and only in particular context a second negation adverb can occur, but in order to express a particular form of topicalization on the negation. See e.g. *Io non mangio mele* (I don't eat apples) and *Io non ho nemmeno mangiato la mela* (I have not even eaten the apple). In the implementation of SimpleNLG-IT we modified therefore that made for French, by considering *non* instead of *ne* and by allowing the presence of a *negation_auxiliary* when the user want (instead of the adverb *pas*).

2.2 Morphonology

In this section we present the issues addressed for making the generated linguistic expression compliant with the morphonological tenets of Italian, like e.g. elision, preposition-article contraction and the fusion of clitics with other words.

2.2.1 Article elision

Elision affects all the Italian articles that precede nouns and adjectives beginning with a vowel. Two simple examples are: (1) ***l'uomo*** (the man) = *lo* [Definite Article Masculine Singular] + *uomo* [Common Noun Masculine Singular] (2) ***un'****interessante proposta* (an interesting proposal) = *una* [Undefined Article Feminine Singular] + *interessante* [Qualitative Adjective Feminine Singular] + *proposta* [Common Noun Feminine Singular]. We adapted with specific rules the morphonological processor introduced in SimpleNLG-EnFr1.1 to manage these cases.

2.2.2 Preposition contraction

Similar to French and Brazilian Portuguese (de Oliveira and Sripada, 2014), Italian provides a morphophonological mechanism to contract the articles and the prepositions which are associated with them in prepositional articles. Among the ten Italian proper prepositions (*di* (of), *a* (to), *da* (from), *in* (in), *con* (with), *su* (on), *per* (for), *tra* (among), *fra* (among)) only three do not contract with the article (i.e. per, tra and fra). For instance, *la casa* ***della*** *zia* (the house of-the aunt) = *la* [Definite Article Feminine Singular] + *casa* [common noun feminine singular] + ***della*** [*di* [preposition] + *la* [definite article feminine singular]] + *zia* [common noun feminine Singular]. Also for this morphophonological phenomenon we added some specific rules in the processor.

2.2.3 Clitics

Clitics are pronouns that in particular cases in Italian can be included in the verb form, like in the following example: *Dam***mi** *la mela* (Give-**me** the apple). More complex forms of clitic-fusion are possible, e.g. *Dam***mela** (Give-**me-it**). However, considering that in most of cases the form with the clitic separated from the verb is tolerated[4], in this phase of the project we decided to simplify clitic morphonology management by applying fusion with the verb only to the pronoun that play direct-object role: if there are other pronouns they are managed by us-

[4] See the distinction between *strong* and *weak pronouns* in (Patota, 2006).

ing prepositions. So, SimpleNLG-IT will generate for the first example above the form *Dai a me la mela* (Give me the apple), where the prepositional phrase *a me* (to me) semantically and pragmatically corresponds to the clitic *-mi*, while the second example will be *Dalla a me* (Give-it to me) where the direct objet clitic pronoun *la* (it [feminine singular]) is fused with the verb but the indirect object (*a me*) is separated.

3 The SimpleNLG-IT lexicon

Each lexicon can be split in two major classes: open and closed classes. The closed class, that is usually composed by function words (i.e. prepositions, determiners, conjunctions, pronouns, etc.) is one to which new words are very rarely added. In contrast, the open class, that is usually composed by lexical words (i.e. nouns, verbs, adjectives, adverbs), is one that accepts the addition of new words. We adopted the same strategy of (Vaudry and Lapalme, 2013): we built by hand the closed part of the Italian lexicon and we built automatically the open part by using available resources.

Additionally, even though, if several lexical corpora are available for Italian, as the detailed map of the Italian NLP resources produces within the PARLI project shows[5], unfortunately most of them are designed to represent lexical semantics rather than morphosyntactic relations. This makes them not adequate for the sake of our task. In order to build the open class of the Italian lexicon, which is suitable for SimpleNLG-IT, we need both a large coverage and a detailed account of morphological irregularities, also considering their high frequency in Italian. Moreover, in order to have good time execution performance in the realiser (cf. (de Oliveira and Sripada, 2014)), a trade-off between the size of the lexicon and its usability for our task must be achieved, which consists in assuming a form of word classification where fundamental Italian words are distinguished from the less-fundamental ones. In order to build a so designed lexicon, we decided to merge the information represented in three existing resources for Italian, namely *Morph-it!* (Zanchetta and Baroni, 2005), the *Vocabolario di base della lingua italiana* (De Mauro, 1985) and, for a specific issue, Wikipedia[6]. The difference between them can be referred to both the reasons for which the authors developed them and the adopted methodology and approach. This makes these resources especially useful for us, since they provide information relevant for SimpleNLG-IT which are the same as observed in different perspective, or complementing each other.

The dataset of the Morph-it! project consists of a lexicon organized according to the inflected word forms, with associated lemmas and morphological features (Zanchetta and Baroni, 2005). The lexicon is provided by the authors as a text file where the values of the information about each lexical entry are simply separated by a tab key. It is in practice an alphabetically ordered list of triples form-lemma-features. An example of the annotation for the form *corsi* (*ran*) is:

```
corsi correre-VER:ind past+1+s
```

where the features are PoS (VERb), mood of the verb (indicative), tense (past), person (1), and the number (singular). The last released version of Morph-it! (v.48, 2009-02-23) contains $505,074$ different forms corresponding to $35,056$ lemmas. It has been realized starting from a large newspaper corpus, nevertheless it is not balanced and a small number of also very common Italian words are not included in the lexicon, e.g. *sposa* (bride), *ovest* (west) or *aceto* (vinegar). Morph-IT! represents extensionally the Italian language by listing all the morphological inflections, i.e. adjective, verbs, nouns inflections are represented as a list rather than by using morphological rules. As a consequence the lexicon is huge and using the whole Morph-IT! in SimpleNLG-IT would cause time complexity problem.

The second main resource we exploited for populating the SimpleNLG-IT lexicon is the "Vocabolario di base della lingua italiana" (VdB-IT henceforth), a collection of $7,000$ words created by the linguist Tullio De Mauro and his team (De Mauro, 1985)[7]. The development of this vocabulary has been mainly driven by the distinction between the

[5] http://parli.di.unito.it/resources_en.html

[6] https://it.wikipedia.org

[7] The second edition of the vocabulary has been announced (Chiari and De Mauro, 2014) and it is going to be released (p.c.).

```
foreach adverb $\in$ Morph-IT! $\cap$ VdB-IT do
    Add the adverb in normal form into $\mathcal{L}$
end
foreach adjective $\in$ Morph-IT! $\cap$ VdB-IT do
    Add the adjective in normal form (masculine-singular) and
    in feminine-singular, masculine-plural, feminine-plural
    forms, into $\mathcal{L}$
end
foreach noun $\in$ Morph-IT! $\cap$ VdB-IT do
    Add the noun in normal form (singular), the plural form, and
    the gender into $\mathcal{L}$
end
foreach verb $\in$ Morph-IT! $\cap$ VdB-IT do
    if the verb is irregular then
        Add into $\mathcal{L}$ all the inflections for the indicativo
        presente, congiuntivo presente, futuro semplice,
        condizionale, imperfetto, participio passato, passato
        remoto
    else
        if the verb is reflexive then
            Set active the reflexive feature in the lexicon
        end
        if the verb is incoativo then
            Set active the incoativo feature in the lexicon
        end
        Add the verb in normal form into $\mathcal{L}$
    end
end
```

Algorithm 1: The algorithm for building the lexicon $\mathcal{L}$

most frequent words (around $5,000$) and the most *familiar* words (around $2,000$). VdB-IT is therefore organized in the following three sections:

- the *vocabolario fondamentale* (fundamental vocabulary), which contains $2,000$ words featured by the highest frequency into a balanced corpus of Italian texts (composed of novels, movie and theater scripts, newspapers, basic scholastic books); *amore* (love), *lavoro* (work), *pane* (bread) are in this section.

- the *vocabolario di alto uso* (vocabulary of high usage), which includes other $2,937$ words with high frequency; *ala* (wing), *seta* (silk), *toro* (bull) are in this section

- the *vocabolario di alta disponibilità* (vocabulary of high availability), is composed of $1,753$ words not often used in written language, but featured by a high frequency in spoken language, which are indeed perceived as especially familiar by native speakers; *aglio* (garlic), *cascata* (waterfall), *passeggero* (passenger) are in this section.

This resource helps us in addressing the issues related to the comprehensibility and readability of the generated texts in the SimpleNLG-IT project: indeed by using only words from the *vocabolario fondamentale* we can be confident that we are generating outputs that will be considered as comprehensible for at least 66% of the Italian speakers (De Mauro, 1985).

PoS	Number	%
Adverb	146	2
Adjective	1333	19
Noun	4092	58
Verb	1451	21
(Irregular)	(283)	(4)
Total	7022	100

Table 2: Number of elements for the open categories in the SimpleNLG-IT lexicon.

VdB-IT helped us to limit the size of the lexicon but does not provide information about verb behavior. We need instead to distinguish regular verbs, that are inflected by using rules extracted from the reference grammar, from the irregular ones. The reference grammar reports a partial list of the principal Italian irregular verbs, but we decided to use the larger list of verbs reported in Wikipedia[8]. Another linguistic distinction for Italian verbs reported in Wikipedia[9] has been exploited in the lexicon: the *incoativi* verbs have a special behavior in the present time and need to be marked in the lexicon. In Algorithm 1 we reported the algorithm for the creation of the SimpleNLG-IT lexicon and in Table 2 we reported some statistics about its composition.

4 SimpleNLG-IT Evaluation

NLG systems can be evaluated by using controlled as well as real world examples: the former examples can be exploited in evaluating specific features of the system, while the latter ones for testing the usability of the system in an application context. In order to provide a first but accurate evaluation of SimpleNLG-IT, we decided to apply both strategies. First, we test the system in the generation of a number of sentences obtained from

[8] `https://it.wikipedia.org/wiki/Verbi_irregolari_italiani`

[9] `https://it.wikipedia.org/wiki/Verbi_incoativi`

SimpleNLG-ENFr1.1. Second, we considered 20 sentences from the Italian section of the Universal Dependency Treebank (Nivre et al., 2016).

We first tested SimpleNLG-IT by running a set of Junit Tests on 96 sentences extracted and adapted from the reference grammar book and from SimpleNLG-EnFr1.1 JUnit Tests. The tests cover different sections of the Italian grammar: adjectives order, different types of sentences (relative, interrogative, coordinated, passive), verbs conjugation, clitics, etc. are analyzed. For this test, the loading into the memory of the lexicon took $1,433$ ms and the test bundle run finished in $3,145$ ms on a computer equipped with 8GB and i7 processor: all the test are passed by SimpleNLG-IT.

In the second evaluation, we wanted to test if SimpleNLG-IT is able to realize sentences from real world. The Universal Dependency Treebank (UD) is a recent project that aims to "create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework" (Nivre et al., 2016). UD released freely available treebanks for 33 languages (in this work, version 1.2). Each UD treebank is split in three sections, *train*, *dev* and *test*, which can be exploited in the evaluation of NLP/NLG systems. Indeed, for the evaluation of the SimpleNLG-it we used the test section of the Italian UD treebank (UD-IT-test). We chose 10 declarative sentences and 10 interrogative sentences, which have length up to ten words, from UD-IT-test. In Table 3 we report the sentences employed. We tried to generate each one of these sentences in SimpleNLG-IT but, since the system can generate canned text, we need to specify a number of *rules* that we respect in order to convert the dependency structure of the sentences into the SimpleNLG input structure: (i) We build a SimpleNLG input *isomorphic* to the gold dependency tree. So, we use the corresponding functions for subject, object, complement, passive verbs etc. (ii) We do not use canned texts and we do not provide information about word order. So we do not use the `insertPreModifier` and `insertPostModifier` functions. (iii) We do not provide information about genre and number for words in the lexicon. (iiii) We do not account for the punctuation inside the sentence.

We obtained very different results for declarative sentences and for interrogative sentences[10]. For declarative sentences we have: two realized sentences are identical to the gold (6, 7); four realized sentences are different only in the word order respect to the gold (1, 3, 8, 10); two realized sentences are different only for clitics respect to the gold (2, 4); one realized sentence is different respect to the gold since the verb is not present in the lexicon (9); one realized sentence is different respect to the gold since the a verb is not treated as irregular (5). In contrast, in interrogative sentences we have more problematic cases: one realized sentence is identical to the gold (19); two realized sentences are different only in the word order respect to the gold (14, 16); one realized sentence is different respect to the gold since the verb is not present in the lexicon (11); six realized sentences are different respect to the gold since the SimpleNLG is not able to apply a WH-question to the specific argument (12, 13, 15, 17, 18, 20), i.e. the realiser is not able to produce HOW-MANY, WHAT, WHICH questions on the object or complements. Finally, we note that most word order errors are caused by the SVO order that is adopted in SimpleNLG-IT. Indeed, the sentences 1, 3, 8, 10, 14, 16 are grammatical but the gold sentences have a different topic-focus information structure represented with a different word order.

5 Conclusions and future work

In this paper we presented the first version of SimpleNLG-IT, a realisation engine for Italian. We introduced with respect to previous implementations a number of new features to account for the morphological and syntactical peculiarities of Italian. We developed a new schema for encoding the Italian verb tense system and a new lexicon by merging two different lexical resources. We performed a first evaluation of the system based on both controlled and real word sentences.

In future work we intend to expand SimpleNLG-IT by using information from UD-IT treebank. In particular, we want to exploit the syntactic information contained in the treebank in order: (1) to decide the correct auxiliary verb to use in order to form complex verb tense, (2) the word order of some adjectives. Indeed, both such notions cannot be ac-

[10] Henceforth the numbers in parentheses refer to Table 3

ID	Gold sentence	Realized sentence
1	*Chiedi al computer il tuo menù.* (Ask to the computer your menu.)	Chiedi il tuo menù al computer.
2	*Dimmi dove si trova la compagnia DuPont.* (Tell me where the DuPont company is.)	Dici a me dove la compagnia Du Pont si trova.
3	*È stato concordato un pacchetto di riforme.* (It was arranged a reform package.)	Un pacchetto di riforme è stato concordato.
4	*Lui le regalò un porcellino salvadanaio.* He gave her a piggy bank.	Egli regalò a lei un porcellino salvadanaio.
5	*È successo un quarto d'ora fa.* (It happened fifteen minutes ago.)	Ha successo un quarto d'ora fa.
6	*Mai nessuna azzurra aveva conquistato un titolo iridato.* (Never any Italian athletes had won a world title.)	Mai nessuna azzurra aveva conquistato un titolo iridato.
7	*L'espropriazione è realizzata attraverso un atto amministrativo;* (The expropriation is carried out through an administrative act;)	L'espropriazione è realizzata attraverso un atto amministrativo;
8	*Non ho preclusioni ideologiche, spiega.* (I have no ideological barriers, he explains.)	Spiega non ho preclusioni ideologice.
9	*Ogni fosso interposto tra due fondi si presume comune.* (Each ditch interposed between two funds is assumed to be shared.)	Ogni fosso interporre tra due fondi si presume comune.
10	*L'insieme di tutte queste operazioni viene chiamato stigliatura.* (The set of all these operations is called decortication.)	L'insieme di queste operazioni tutte è chiamato stigliatura.
11	*In che modo le Hawaii divennero uno stato?* (How did Hawaii become a state?)	Come le Hawai divenirono uno stato?
12	*Quante fossette ha una pallina regolamentare da golf?* (How many dimples does a regular golf ball have?)	-
13	*Da quante repubbliche era composta l'Unione Sovietica?* (How many republics did compose the USSR?)	-
14	*E i soldi delle piramidi dove sono finiti?* (And where did the money of the pyramids go?)	Dove i soldi delle piramidi sono finiti?
15	*Che cosa ha influenzato l'effetto Tequila?* (What did influence the Tequila effect?)	-
16	*Quanto si stima che costeranno le stazioni spaziali internazionali?* (How much is estimated that will cost the international space stations?)	Quanto si stima che le stazioni spaziali internazionali costeranno?
17	*Quali paesi ha visitato la first lady Hillary Clinton?* (Which countries did the first lady Hillary Clinton visit?)	-
18	*In quale giorno avvenne l'attacco a Pearl Harbor?* (Which is the date when Pearl Harbor was attacked?)	-
19	*Quando Panama si vide restituire il Canale di Panama?* (When did Panama see to return back the Panama Canal?)	Quando Panama si vide restituire il Canale di Panama?
20	*Da quale animale si ricava il veal?* (From which animal do you get the veal?)	-

Table 3: Ten sentences from the UD-IT-TEST: 1-10 are declarative sentences and 11-20 are interrogative sentences.

counted by using rules from grammar books but they need an empirical approach. Finally, in order to have a larger set of tests, we want to develop an algorithm for automatically convert dependency tree of UD-IT in SimpleNLG-IT input. In this way, we can use the whole test section of the treebank as benchmark.

References

Marcel Bollmann. 2011. Adapting SimpleNLG to German. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138, Nancy, France, September. Association for Computational Linguistics.

Isabella Chiari and Tullio De Mauro. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *1th Italian Conference on Computational Linguistics (CLiC-it)*, volume 1, pages 93–97. Pisa University Press, December.

Tullio De Mauro. 1985. *Guida all'uso delle parole*. Libri di base. Editori Riuniti.

Rodrigo de Oliveira and Somayajulu Sripada. 2014. Adapting SimpleNLG for Brazilian Portuguese realisation. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 93–94, Philadelphia, Pennsylvania, U.S.A., June. Association for Computational Linguistics.

Sasi Raja Sekhar Dokkara, Suresh Verma Penumathsa, and Somayajulu Gowri Sripada. 2015. A Simple Surface Realization Engine for Telugu. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 1–8, Brighton, UK, September. Association for Computational Linguistics.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece, March. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC 2016), Paris, France, may. European Language Resources Association (ELRA).

Giuseppe Patota. 2006. *Grammatica di riferimento dell'italiano contemporaneo*. Guide linguistiche. Garzanti Linguistica.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Manuela Sanguinetti, Cristina Bosco, and Leonardo Lesmo. 2013. Dependency and constituency in translation shift analysis. In *Proceedings of the 2nd Conference on Dependency Linguistics (DepLing)*, pages 282–291, Prague (Czech Republic). Charles University in Prague, Matfyzpress.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting SimpleNLG for Bilingual English-French Realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187, Sofia, Bulgaria, August. Association for Computational Linguistics.

Micheal White. 2006. Efficient realization of coordinate structures in combinatory categorial grammar. *Research on Language and Computation*, 2006(4(1)):39—75.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, 1(1).