VenPro: a morphological analyzer for Venetan

Sara Tonelli[1], Emanuele Pianta[1], Rodolfo Delmonte[2], Michele Brunelli[2]

[1]Fondazione Bruno Kessler, Trento, [2]Dept. of Language Sciences, University of Venice,
{satonelli,pianta}@fbk.eu ; {delmont,michele.brunelli}@unive.it

Abstract

This document reports the process of extending MorphoPro for Venetan, a lesser-used language spoken in the Nort-Eastern part of Italy. MorphoPro is the morphological component of TextPro, a suite of tools oriented towards a number of NLP tasks. In order to extend this component to Venetan, we developed a declarative representation of the morphological knowledge necessary to analyze and synthesize Venetan words. This task was challenging for several reasons, which are common to a number of lesser-used languages: although Venetan is widely used as an oral language in everyday life, its written usage is very limited; efforts for defining a standard orthography and grammar are very recent and not well established; despite recent attempts to propose a unified orthography, no Venetan standard is widely used. Besides, there are different geographical varieties and it is strongly influenced by Italian.

1. Introduction

In this document we illustrate a project for the creation of a morphological analyzer / synthesizer for a lesser-used language[1] of Italy, namely *Venetan*, and we describe the process of extending MorphoPro (Pianta et al., 2008), a modular application for morphological analysis, to this extent. This new implementation is part of the STILVEN project (Delmonte et al., 2009), and the creation of an NLP architecture for Venetan is seen as a preliminary step to the development of a machine translation system from Venetan to English. This work is motivated also by the increasing interest of the NLP community in resource-poor languages, with dedicated workshops and important publications devoted to them (see for example the last IJCNLP Workshop on NLP for Less Privileged Languages, `http://ltrc.iiit.ac.in/nlp-lpl-08/`). Anyhow, attention was mainly focused on "minority" or "lesser-used" languages that have a somewhat official status, for example Galician (González et al., 2008), Catalan (Torres et al., 2002) and Basque (Alegria et al., 2002). These languages are officially adopted in some regions, are taught in schools, they are used in written texts and sometimes on TV, and there are Institutions devoted to their protection. On the contrary, the so-called Italian dialects are mainly spoken and, apart from some exceptions (for example the *Ladin* language in Trentino (Bortolotti and Rasom, 2005)), haven't deserved much attention from the NLP community. Italy has developed a variety of local dialects mainly influenced by Latin and by the subsequent substrata influence of barbaric invasions. Differently from other nations in Europe, in Italy the Latin from which current dialects stem is the one that was spoken before the time of the greatest expansion of the Roman Empire. There are at least 20 different main dialects including Sardinian, but in fact, every dialect has at least 2 varieties. Thus there is a total approximative number of 50 dialect varieties spoken in Italy. No other nation in Europe or elsewhere possesses such a rich inventory of languages[2]. For this reason, we believe it is worth studying them from a computational point of view and developing NLP tools for their analysis can be seen as a step towards their valorization and preservation. Furthermore, Venetan and Italian are both Romance languages that show a high degree of morphosyntactic similarity, so that we can take advantage of existing NLP tools for Italian and adapt them to Venetan instead of creating them from scratch. A similar approach was successfully adopted also for Zulu and Xhosa, two Nguni languages spoken in South Africa (Pretorius and Bosch, 2009).

In the following section we briefly describe the history of Venetan and introduce the STILVEN project. Then, in Section 3. we give an overview of the main characteristics of Venetan morphology compared to Italian. In Section 4., we present the TextPro suite, specifically the MorphoPro module for morphological analysis and we describes the main part of our contribution, which is the development of rules for inflection morphology analysis of Venetan. In Section 5. we discuss the ongoing evaluation of the morphological analysis and finally we draw some conclusions in Section 6.

2. *Venetan* and the STILVEN project

STILVEN is a project funded by the Venetan Region in Italy which was started in February 2008 and that involves the Laboratory of Computational Linguistics at the University of Venezia and FBK-Irst in Trento. The aim of the project is the creation of a computational infrastructure for the translation of Venetan into English.

Venetan[3] has been the official language of the Veneto Republic for as long as 8 centuries, up to the moment in which

[1]There is an ongoing debate about the definition of Venetan as a language or as a dialect. We use the term *lesser-used language* instead of *dialect* for the sake of clarity, but we don't intend to make any assessment about the current debate, which is influenced mainly by political considerations and focuses on sociopolitical aspects rather than on linguistic issues.

[2]For a map of Italian dialects and several working papers, see `http://asis-cnr.unipd.it/`

[3]We follow (Maiden and Parry, 1997) and use the word *Venetan* to refer to the whole language spoken under different varieties. The word *Venetian*, which is sometimes used in the literature as a synonym of Venetan, should be reserved only to one variety, namely the one spoken in the city of Venice.

the Republic was included in the newborn Italian nation at the end of the XIXth century. Since then, Venetan has been slowly abandoned in favour of Italian. In spite of that, Venetan has a much wider usage than other Italian dialects and it is commonly spoken in most working places, families and in the social life of the Veneto region. Besides, it is used also in other regions, such as Trentino, Friuli Venezia Giulia, Istria and some towns of Dalmatia, so that the number of native speakers amounts to more than two million. In the Veneto region, Venetan proficiency by local speakers has been lately assessed as reaching 75% of the population, even if it is no longer a language used by administration and other official institutions. A small community of Venetan speakers is very active on the Internet, contributing to the diffusion of this language through several web-sites including a version of Wikipedia in Venetan (`http://vec.wikipedia.org/wiki/Vèneto`). As for written texts, there is a literary tradition of Venetan, especially of its varieties (for example, the works in Venetian by the playwright Carlo Goldoni or the poems by Biagio Marin in the Grado dialect), but the unified Venetan has just few attestations as everyday written language.

The implementation of NLP systems for lesser-used languages like Venetan is particularly challenging for three reasons: first, the number of *varieties*, second the fact that it has no estabilished *orthography* and third the influence of Italian. Assuming that a variety must show lexical, phonological and structural differences that enable the hearers to understand the speaker's provenance, linguists have identified at least 4 varieties of Venetan. For this reason, Venetan is actually a 'diasystem', where speakers use their own variety and manage to understand each other. The low level of standardization of the written form is a common problem among Italian dialects, even if there have been many attempts of orthography unification and normalization. The relevance of this problem and the interest of the Regional institutions in the normalization issue is proved by the fact that the Regional Government has released an official document called *Manual of Venetian Orthography* (Giunta Regionale del Veneto, 1995), written by a Scientific Committee, with the description of the most common orthographic variations of Venetan. Finally, the strong influence of Italian on dialect speakers affects the quality of spoken Venetan, since utterances often include Italian loans.

3. Peculiarities of Venetan Morphology

Venetan is a romance language and as such it shares a number of properties with other Latin-derived languages. For instance, the Venetan nominal system is based on two genders (masculine/feminine) and two numbers. The form *vecio* is used as an adjective ('old') or noun ('old man') in the whole Venetan area and its inflection (1a, 2a) mirrors that of its Italian counterpart *vecchio* (1b, 2b).

1a)	vecio	veci	(Venetan)
1b)	vecchio	vecchi	(Italian)
	old-m.sg.	*old-m.pl.*	
	'old man'	'old men'	

2a)	vecia	vecie	(Venetan)
2b)	vecchia	vecchie	(Italian)
	old-f.sg.	*old-f.pl.*	
	'old woman'	'old women'	

The verbal system of Venetan has a quite large number of mood, tense and person affixes, very similar to those encountered in Italian or Spanish. A comparative example is shown in Table 1. The Venetan verb *cantar* (to sing) has exactly the same root as its Italian and Spanish translation and also the form of the first person singular of the present indicative tense is identical.

The pronominal system of Venetan makes a distinction between strong and clitic pronouns, as observed in other romance languages. See for instance the forms of the object pronoun in the 1^{st} singular person (examples 4) and 1^{st} plural person (examples 5). Even if some forms are "inverted" with respect to their Italian counterparts (*mi/me* in examples 4), the behaviour of strong and clitic pronouns in Venetan and in Italian is very similar.

	(to) me-strong	*me-clit*	
4a)	(A) mi	me	(Venetan)
4b)	(A) me	mi	(Italian)

	(to) us-strong.m/f	*us-clit*	
5a)	(A) noaltri/noaltre	ne	(Venetan)
5b)	(A) noi	ci	(Italian)

Nevertheless, Venetan shows some prominent peculiarities. In Venetan, gender marking on pronouns is more pervasive than in Italian since distinct masculine and feminine forms are usually employed not only in the 3rd person, but also in the 1st plural person and 2nd plural person. See, for instance, the forms *noaltri* (m.plur) and *noaltre* (f.plur) in the example 5a). Case marking, in contrast, has completely disappeared on strong pronouns. One and the same form acts as both subject and object as, for example in: *mi* 'I/me', *ti* 'you (thou/thee)', *noaltri* 'we/us (masc.)', *łóre* 'they/them (fem.)' and so on. Case, however, is marked on clitics, along with gender and number. In Venetan, clitics can be employed as direct objects, indirect objects and also subjects. In Table 2, for example, the sentence 6a) contains three clitics in sequence, respectively in subject (*el*), indirect object (*ghe*) and direct object (*ło*) position. The Italian equivalent has only one clitic, *glielo*, bearing the function of indirect and direct object. Note also the Venetan translation of the Italian clitic *lo*, which is *ło* in our example but is often transcribed as *£o* and also as *lo*.

Another important property of Venetan is that analytic, periphrastic constructions are usually preferred over other forms. For instance, the simple past (Ital. *passato remoto*) has completely disappeared and has been replaced by the present perfect, composed by an auxiliary verb and a past participle. This means that the morphological analyzer will have to generate and recognize less verbal tenses for Venetan that for Italian.

There is a certain tendency in Venetan to prefer prenominal and preverbal elements (articles, subject clitics) over endings to encode gender, number and person features. Due to sincretism and processes of morphological reduction,

3a)	canto	cantè	cantàvimo	(Ven.)
3b)	canto	cantate	cantavamo	(Ita.)
3c)	canto	cantais	cantàbamos	(Spa.)
	sing-1^{st}*sg.pres.ind*	*sing-*2^{nd}*pl.pres.ind*	*sing-*1^{st}*pl.imperf.ind*	
	'I sing'	'You sing'	'We used to sing'	

Table 1: Example of verb morphology in Venetan, Italian and Spanish

6a)	El	ghe	ło	porta	doman	(Venetan)
	subj.cl.3^{rd}*m.sg*	*ind.obj.cl.*3^{rd}	*obj.cl.*3^{rd}*m.sg*	bring	tomorrow	
6b)		Glielo		porta	domani	(Italian)
		ind.obj.cl.3^{rd} + *obj.cl.*3^{rd}*m.sg*		bring	tomorrow	
	'He brings it/him to him/her/them tomorrow'					

Table 2: Pronoun behaviour in an Italian and Venetan sentence

some endings are ambiguous or reduced to zero. The extent of these phenomena depends on the variety, as shown in examples (12a-c): while Venice Venetan presents two different endings for the singular and plural form of *fero* (iron), i.e. *-o* and *-i*, the Treviso variety presents a zero ending in the singular form. As for the Belluno Venetan, both inflectional endings are reduced to zero, which neutralizes the singular/plural opposition of the noun.

12a)	fero	feri	(Venice Venetan)
12b)	fer	feri	(Treviso Venetan)
12c)	fer	fer	(Belluno Venetan)
	iron-m.sg.	*iron-m.pl.*	
	'iron'	'irons'	

Articles and subject clitics, in contrast are more stable throughout the language. For instance, the masculine definite articles are *el* (m.sg.) and *i* (m.pl.) in the vast majority of the Venetan area and grant the encoding of the relevant features. The plural article *i* marks gender and number even when there is no ending on the noun. For instance, the complete plural form of 12c) is *i fer*, (the irons/the iron tools). This means that the morphological analyzer in the recognition step must be as flexible as possible to cope with all existing versions of a form. Such flexibility is not required for the analysis of Italian morphology.

The phenomenon of zero endings concerns also verbal forms. Some endings may be omitted, while subject clitics grant the encoding of person, number and gender features. Italian, in contrast, marks person and number by means of endings which cannot be omitted. Some forms with zero ending are commonly used (or accepted) in a big part of the Venetan area, but the pattern of variation is complicated by the fact that sincretism has worked in different ways in different varieties (Marcato and Ursini, 1998).

Another particular feature of Venetan is the presence of *i*-metaphony in some varieties. In particular, the presence of a final *-i* often induces a change in the root vowels of noun, adjectives and verbs. See for example the forms *fiori/fiuri* ('flowers'), *tenpi/tinpi* ('times') and *meti/miti* ('you put'). Despite being sometimes reported as obsolete or on its way to disappear, this phenomenon seems productive enough to yield forms like *te parchegi/ te parchigi* ('you park the car'). This fact, together with the presence of ambigous or zero endings, contribute to a yet higher variation in the inflectional paradigm of Venetan nouns and verbs. Nouns and verbs may come to have three different forms, even if they belong to the regular inflectional class, because the rule operates slightly different in the different varieties. Such a variation, instead, is not observed in Italian. Once again, the point of reference is provided by preverbal and prenominal elements. Whereas nouns and verbs do not show a homgeneous behavoiur per se, articles and clitics can provide the necessary information to the algorithm in a way that is regular for the whole Venetan area.

4. TextPro and MorphoPro

TextPro (Pianta et al., 2008) is a suite of tools designed for a number of NLP tasks such as Web page cleaning, tokenization, sentence splitting, morphological analysis, PoS-tagging, lemmatization, multiword recognition, chunking and named-entity recognition. The suite has been designed so as to integrate and reuse state of the art NLP components and is freely available for research purposes (`http://textpro.fbk.eu/`). The TextPro architecture is based on a pipeline of processors: each processor accepts data from an initial input or from the output of a previous processor, executes a specific task, and sends the resulting data to the next stage, or to the output of the pipeline. Pipelines of processors are widely used in building NLP applications, mainly due to their simplicity and flexibility.

MorphoPro is a morphological analyzer / synthesizer comprising a development environment, implemented in Prolog, and a run-time version implemented in C++.

The development environment allows for defining a declarative representation of the knowledge needed to analyze and synthesize a given language. MorphoPro rules are first defined as a context-free grammar (CFG) with attributes. A module for *morphological adjustment* applied by the analyzer after a rule has generated a form is further inspired by two-level morphology (Koskenniemi, 1983). With the development environment it is possible to create a two-column table containing on each row a word form and its morphological analysis. The table is then compiled in a

very compact and efficient Finite State Automaton (FSA), which is actually used by the run-time version of MorphoPro. For each input word, the tool delivers all possible morphological analyses, which are represented as sequences of features separated by "+". The first two features of any analysis are always the lemma and lexical category, followed by a variable list of other features such as gender, number, etc. For example, the noun *vinçidóri* (winners) is analyzed as follows: *vinçidóri vinçidór+n+m+plur*
In the development phase, the expressiveness of CFG is particularly suited to create a convenient and concise representation of morphological information. In the run-time version, instead, the FSA has proved to be less expressive but more efficient.

4.1. Morphological analysis of Venetan

The approach required for the development of NLP tools for Venetan is different from the normal procedure followed for standardized languages such as English, German or Italian. Statistical approaches require large annotated corpora for training, which is impossible for Venetan because it presents different varieties and has scarce written attestations. On the other hand, the traditional rule-based systems rely on the idea that a language has a standard form, and cope with one specific, well-defined and well-described language variety. In our case, we just had a small reference grammar, the *General Grammar Book of the Venetan Language and its Varieties* (Brunelli, 2007), which is not extensive, and we often had to ask Venetan speakers for further rules and exceptions. So we basically adopted a "pragmatic" rule-based approach, trying to include all orthographical and morphological variations, instead of a "normative" approach, which would have forced us to select one variety and to exclude for example words coming from Italian. This choice is motivated also by the final goal of the STILVEN project: since we want to develop a Venetan-English machine translation system, we have to account for the different varieties of Venetan to deliver an accurate analysis of the source language. If the translation direction was from English to Venetan, it would have been enough to choose one version for generation.

4.2. Morphological information

In order to produce a morphological analysis of Venetan, we first defined a feature-based context-free grammar whose rules define how roots combine with affixes. Each root is assigned a lexical class (e.g. verb, noun, adjective, etc.) and a morphological class which determines the affixes it can combine with. For instance, the morphological class of verbal roots is determined by the conjugation of the verb and the transitivity/intransitivity/reflexivity feature. The same root can be present in several entries with different lexical categories and / or morphological classes. In this way, we can account for different uses of the same verb (for example *alzarse* (refl., "stand up") vs. *alzar* (tr., "raise")) but also for orthographical variations of the same sound (*alzar* and *alsar*, *destrùxer* and *destrùzer*, "destroy") and for the different Venetan varieties (*bévar* and *béver*, "drink", following two different conjugations). Irregularity, such as irregular plurals, is indicated by a feature as well. Affixes, instead, contribute the *inflectional* features of the final word. For example, all verbs belonging to the first regular conjugation class are associated to a set of affixes expressing features such as mood, tense, person and number. By combining roots with compatible affixes, we generated all pairs of form + morphological analysis that comply with the grammar. Then the list was compiled as a Finite State Automaton.
Even if an accepted standard for Venetan does not exist, it is important to cover all possible variations of a word without overgenerating. To this purpose, it is possible to mark a particular set of rules in MorphoPro as belonging to a *pseudo-standard* and others as a *substandard*. In this way, two different language models can be produced: if we activate only the *pseudo-standard* rules, a smaller set of forms is produced, which is particularly suited to the generation task so as to avoid variations of the same form. If we activate all rules, an extended language model is created. Since it would cover all possible variants of each form, it is particularly suited to the recognition task.

4.3. Lexica

In order to create the list of roots, we extracted in a semi-automatic way lemmas from different online sources, such as the English-Venetan online dictionary (`http://www.elgalepin.com/`), and the dictionary of Venetan languages and its varieties (Brunelli, 2006). The latter in particular allowed us to include in our resource several local versions of the same lemma. These lexical resources were then merged and we applied some routines to automatically obtain the roots. In particular, starting from the quotation form and the lexical category of the dictionary, we automatically derived the morphological class of the lemma. For example, given a dictionary entry for the word *debolézha* (weakness), we can derive the root *debolézh*, which belongs to the morphological class of nouns with affix "-a" for singular and "-e" for plural. The morphological information required by MorphoPro were eventually added by lexicographers.
The creation of the lexicon is still ongoing and at present it is based on the semi-automatic extraction of roots starting from tokens taken from online texts. Since many Venetan speakers largely use Italian loans in their utterances, we can exploit the modular structure of MorphoPro to keep also the Italian lexicon as a backup in case a word is not recognized by the Venetan analyzer.
At the time of writing, the lexicon includes 3,900 verbal roots, 19,000 nominal roots and 2,400 adjectival roots. Even if the lexicon dimensions are not comparable to the lexica available for well-estabilished languages, this represents the largest existing lexical resource for Venetan, and the first attempt to cope with different orthographical variations and Italian loans.

5. Evaluation

Evaluation is still ongoing. During the development of the morphological analyzer, the forms generated were verified for every morphological class and also a small development text was employed for form recognition and analysis. At present, the main issue of evaluating texts taken from

different sources is the recognition of orthographical variations. In particular, texts from websites tend to contain much less accented forms as those we encoded in the lexicon and in general their orthography is less accurate. We are developing rules for carrying out a quick check of accented forms and allow for guessing / recognition also of forms with missing accents.

6. Conclusions and future work

In this work we have presented VenPro, the only existing morphological analyzer for Venetan. Since VenPro has been extended starting from MorphoPro, a morphological tool available for English and Italian, we could take advantage of the existing framework, and in particular of the similar inflectional system of Venetan and Italian. Even if a standard for Venetan does not exist and the creation of a morphological tool is very challenging due to the different local varieties of this language, we were able to enrich the lexicon with a high number of roots starting from available online resources such as Venetan Wikipedia.

In the future, we will integrate the tool in an NLP infrastructure for automatic translation from Venetan to English. Such system is being developed in the Stilven project with the aim of making the translation tool available through the web and will be first tested by the large community of Venetan-speakers who live beyond the Italian border.

7. References

I. Alegria, M. Aranzabe, N. Ezeiza, and R. Urizar, 2002. *Using Finite State Technology in Natural Language Processing of Basque*, pages 1–12. Lecture Notes in Computer Science 2494. Springer.

E. Bortolotti and S. Rasom. 2005. Il ladino fra polinomia e standardizzazione: l'apporto della linguistica computazionale. In *Proceedings of the Workshop on Lesser-Used Languages and Computational Linguistics*, Bozen, Italy.

Michele Brunelli. 2006. The Dictionary of Venetian Language. Available online at http://www.dizsionario.org/dizsionario.php.

Michele Brunelli. 2007. General Grammar Book of the Venetan Language and its Varieties. Available online at http://www.michelebrunelli.com/mgx_veneto_en.pdf.

Rodolfo Delmonte, Antonella Bristot, Sara Tonelli, and Emanuele Pianta. 2009. English/Veneto Resource Poor Machine Translation with STILVEN. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, Besançon, France.

Giunta Regionale del Veneto. 1995. Manuale di Grafia Veneta. Available online at http://www.veneto.org/gvu/.

Manuel González González, Eduardo Rodríguez Banga, Francisco Campillo Díaz, Francisco Méndez Pazó, Leandro Rodríguez Liñares, and Gonzalo Iglesias Iglesias. 2008. Specific features of the Galician language and implications for speech technology development. *Speech Communication*, 50(11–12):874–887.

K. Koskenniemi. 1983. Two-level morphology: A general computational model for word-form recognition and production. Publication 11, University of Helsinki, Department of General Linguistics.

M. Maiden and M. Parry. 1997. *The dialects of Italy*. Routledge, London / New York.

G. Marcato and F. Ursini. 1998. *Dialetti veneti: grammatica e storia*. Padova Unipress.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro tool suite. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.

Laurette Pretorius and Sonja Bosch. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 96–103.

Marta Torres, Lluís Jardí, Núria Alturo, Lluís Payrató, and F. Xavier Vila, editors. 2002. *Actes de la I Jornada sobre Comunicació Mediatitzada per Ordinador en Català (CMO-Cat)*.