English/Veneto Resource Poor Machine Translation with *STILVEN*

Rodolfo Delmonte, Antonella Bristot, Sara Tonelli, Emanuele Pianta
Università Ca' Foscari - Department of Language Sciences
E-mail:delmont@unive.it

Abstract

The paper reports ongoing work for the implementation of a system for automatic translation from English-to-Veneto and viceversa. The system does not have parallel texts to work on because of the almost inexistence of such manual translations. The project is called STILVEN and is financed by the Regional Authorities of Veneto Region in Italy. After the first year of activities, we managed to produce a prototype which handles Venetian questions that have a structure very close to English. We will present problems related to Veneto, basic ideas, their implementatiion and results obtained.

1. Introduction

STILVEN is a project approved December 2007 which started its activities in February last year. The task was creating a computational infrastructure to be directed to the analysis and translation of Veneto language. Veneto is a dialect nowadays but was the official language of the Veneto Republic for as long as 8 centuries, up to the moment in which the Republic occupied by the French and then by the Austrian became part of newborn Italian nation at the end of the XIXth century. Since then, Veneto has been slowly abandoned in favour of Italian. The same happened all over Italy. Nowadays, depending on the region, Italian speakers can usually master a dialect and the main language. In particular, Veneto speakers show a much wider usage of dialect – their original language – in most working places, in the family and in social life. Veneto proficiency by local speakers has been lately assessed as reaching 75% of the population.

Being no longer a language used by administration and other official institution, Veneto has become a spoken dialect which however has developed a number of varieties: there are at least 4 which can be regarded as such. A variety is considered as such in case it has both lexical, phonological and grammatical/structural differences that make it clear to the hearing of the local speaker its provenance. Veneto dialect is now considered a diasystem where speakers use their own variety and manage to understand each other.

In addition to varieties, a spoken-only language does not have a fixed orthography. So, even though there have been many attempts at unifying the orthography within and without each variety, speakers have difficulties in using such unified paradigms. In addition, the influence of Italian orthography is very strong.

Then the first problems to be coped with are:

- accounting for the varieties
    - in the lexicon
    - in the grammar
    - in the orthography
- accounting for the orthographic variations

As to the first such problems, we have implemented a number of different lexica which refer at the same time to the four main varieties, to Italian and to English. Reference to Italian will be clarified below.

As to the second problem, it has been solved partially – more on this problem below. The currently implemented solution takes into account possible orthographic ambiguities and produces a uniform output to be matched against the main translation lexicon.

1.1. The General Project

A translation system that has to cope with varieties has three main problems to solve:

- lexicon extension including all specialized items present in one variety and not in the others;
- grammatical flexibility that may adjust different structural organization and treat them in a "normalized" fashion;
- orthographic normalization according to a standard.

We said above that the number of varieties that show remarkable differences are four. However, from a more pragmatic point of view, differences may lead to establish a more detailed panorama of the possible varieties as Manlio Cortelazzo has done in his book

which can be downloaded online at http://www.linguaveneta.it/sussidiario.html, which we report here below:

1. Venetian

- **Ciao Bepi, dove ti va?** /Hi Joseph where are you going
- **So drio sercar mia mugèr: ti la ga vista?** /I'm looking for my wife: have you seen her
- **Sì, la go vista ndar drento col fio de to nèssa dal bechèr qua darente.** /Yes, I have seen her go inside the butcher with the son of your niece here on the back

2. Vicentinian

- **Oh, Ada, dove veto? come steto?** /Oh Ada where are you going? how are you
- **Cossa voto ca te diga? né ben né mae.** /What do you what me to tell you? No good no bad
- **Come mai? Proprio ti, che te jéri sempre alegra!** /How comes? Just you, who were always so jolly
- **Xe vèro ma nò se pòe fermar el témpo.** /True, but time cannot be stopped
- **Te ghè razòn. Te lasso perché gò paura che tra pòco scravassa.** /You're right. I leave you because I am afraid that it will rain cats and dogs shortly

3. Rovigoto

- **Ciao, Maria. Elo mina vegnù to marìo?** /Hi Maria. Has your husband by chance come?
- **Nel poe, l'è vizzìn a la vaca, che speta un vedelìn, ma ghe xe chi so fradeo.** /He can't, he's close to the cow, who is bearing a small calf, but his brother is here
- **Valo mina via?** /Is he not going away
- **Bison ch'el vaga.** /It is necessary that he goes

4. Padovan

- **Ciao, Toni, còssa ghèto che te vedo cossì zò?** /Hi Toni what's going on, you look so worried
- **Taxi taxi, compare, che go me mojère che la sta poco ben.** /Don't mention it, friend, I have my wife who is not feeling well
- **Còssa ga£a?** /What has she got
- **Mah, so drio ndare a ciamare el dotore, che sta sora el caegaro.** /Don't know, I'm going to call the doctor who lives over the shoemaker

5. Trevisan

- **Dove vatu, Teresa?** /Where are you going, Teresa
- **Vao crompar do fòje de salata e un pèr de vovi.** /I'm going to buy two leaves of salad and a couple of eggs
- **E chi xélo sto cèo qua?** /And who is this boy here
- **El xé so fradèl de la Maria, la lavandera.** /He is the brother of Maria, the washer

6. Belunese

- **Féu che, toxat?** /what are you doing boys
- **Porton dentro le tòle parché fra poc piove.** /we bring the chairs inside because shortly it will rain
- **Piòvelo? Ma va là. Chi élo sto bòcia? Come te ciàmetu?** /Rain? You must be joking. Who is this guy? What's your name
- **Tonin.**
- **Ve salude! Sani!** /Let's go, bye! See you

7. Veronese

- **Bepi, com'ela che ti si solo? Dov'èle ndè to fiole?** Joseph, how comes you're alone? Where have your daughters gone
- **A crompàr calcossa da magnar.** /To buy something to eat
- **E ci èlo sto bel buteleto?** /And who is this handsome boy
- **L'è Michele, el fiol de me neòda.** /He is Michael, the son of my niece

All seven dialogues deal with similar topics and present both questions and declaratives as answers. Following Cortelazzo, the four varieties should be organized as follows:

- Venetian
- Vicentinian, Rovigoto, Padovan
- Trevisan, Belunese
- Veronese

However if we try to find similarities and differences, it is the second that we find more easily. As to similarities, all varieties apart from Venetian use subject clitic inversion in questions. Lexical differences are many and constitute the main distinguishing element. Vicentino uses "ca" rather than "che", and "vòto" rather than "vùtu" in Treviso and "vùto" in Verona. Veronese is the only one to use palatalization for interrogative pronouns (chi□ci). Belunese is the only variety to allow verb fronting before question word: "Féu che" / Do what, and clitic subject for weather verbs, "Piòvelo". Then we may note other distinguished uses of lexical items:

- céo / boy, is only used by Trevisan
- Sani / See you, is only used by Belunese

As to the remaining differences, they are all understood by the majority of Veneto people.

So basically, this is what the system should also do: allow for varieties and take care of specialization which are mainly lexical. Syntactic peculiarities will be discussed below.

2. Orthographic Normalization

Many languages in the world share the same problem of orthographic variation – Arabic, Chinese, Japanese, Korean etc. -, hence the need to produce a normalization that allows the wordform to be checked against a lexicon where standardized orthography has been used. In our case, lexemes are produced in the lexicon with an official orthography according to the GVU (Unified Veneto Writing) obeying rules formulated some years ago elaborated by linguists and published in the website of Veneto Region.

To make a comparison with Arabic, we see that orthographic variations may arise for a number of reasons, the first of which is certainly the dialectal variation. Then there is the objective problem of rendering some phonemes into a romanized valid corresponding character. As a result, an Arabic name may have hundreds if not thousands different variants in its romanized version. Coming back to Veneto, the problem is not so acute and the solution that can be adopted is the one that is also applied to other languages, that is an orthographic rule-based approach. In other words, due to the small number of variants it is not fit to use a lexicalized approach where all variants are stored after being validated automatically and then manually, on the basis of their frequency of occurrence on the web, for instance. It will then be sufficient to list all cases of orthographic variations occurring in Veneto and then to formulate a corresponding set of rules. These rules coincide with what has been done for Arabic, for instance. In particular, consider the following rules for the recognition of some typical characters. As may be seen, the starting point is the corresponding phoneme, and on the right hand side there is a list of possible graphemes

/dz, ts/	□	d dh t z th
/k/	□	k q c ch
/j/	□	j g dj

where we see that there is one-to-many mapping. In the case of Veneto, /k/ presents the same mapping problems; on the contrary /T/ could correspond to /dz/ or /ts/ which is Veneto is rendered sometimes and only in some variants.

From the two tables below it is also clear that there is no correspondence between Veneto and Italian as far as graphemes are concerned. Not only Veneto lacks geminates, but it uses the same Voiced S sound for a variety of graphemes in Italian corresponding words as shown below.

IT Grapheme	VE	Italian	Veneto	Translation
s	→ x	posa	pòxa	pose
c	→ x	pace	pàxe	peace
gi,ge	→ x	peggio	pèxo	worse
zz	→ x	mezzo	mèxo	half
z	→ x	zero	xèro	zero

Table 1. Italian/Veneto [x] grapheme mismatch

As can be seen, /x/ may correspond to Italian /s/ /tch/, /dg/ /dz/ as far as sounds are concerned, and to [s, c, gi, ge, z, zz] as far as graphemes are concerned. The same happens with Veneto /s/ as shown in Table 2. below, where it may correspond again to /s/ /z/ /tch/ /sc/, and to graphemes [ss, c, zz, z, sci, sce],

IT Grapheme	VE	Italian	Veneto	Translation
ss	→ s	passo	pàso	step
c	→ s	piacere	piasér	please
zz	→ s	pozzo	póso	water wel
z	→ s	pozione	posión	potion
sci, sce	→ s	pesce	pése	fish

Table 2. Italian/Veneto [s] grapheme mismatch

The other remarkable orthographic problem concerns the need to use word stress on E and O to differentiate open vs. closed phoneme. The difference is crucial to characterize minimal pairs which otherwise would not be disambiguated, as we can see from the examples below,

béco [goat]	bèco [beak]
péxo [weight]	pèxo [worse]
bóte [keg]	bòte [strikes]
fóla [crowd]	fòla [lie]
etc…	

So here again the problem lies in the lack of awareness on the side of the native speaker of the need to introduce such diacritics because they don't hear the ambiguity. Normalizing in this case is more complex. The question here is that the meaning changes according to the type of accent chosen. In all of these cases then the two variants need to be present until the translation takes place: at that moment, semantic word disambiguation processes need to be activated in order to select the correct words compatible with the context. To this end we organized specialized vectors of lexical fields, i.e. words related to each of the meanings of the minimal pair. This vector of words will be searched each time semantic disambiguation has to be activated. More on this problem below.

We have been working only at the Veneto-English translation module because it is easier to produce given the much richer lexicon of Veneto when compared to

English. In particular, in a section below we will present preliminary work related to the treatment of interrogatives which show a high structural isomorphism with English. As to the remaining sentence types, we will use Italian as intermediate language onto which build the shallow syntactic representation to use for the Transfer module. In that case, we already took advantage of the Italian-English parallel corpora available online to search for frequent multiwords translation pairs. Eventually, we will produce phrase reordering rules at the level of logical form, in order to recover correct Predicate Argument Structures (PAS) in the target language, English (see [2]).

In addition to the rule-based approach, we are trying to develop a statistical machine translation module based on a small set of parallel corpora we have collected. They are approximately 70,000 tokens and we intend to use it to develop a language model using GIZA and MOSES.

3. Resources and NLP Tools

Very much like what has been done with METIS (see [3]) STILVEN aims at translating free text input by taking advantage of a combination of statistical, pattern-matching and rule-based methods. The following goals and premises were defined for the project:

1. use simple NLP tools and resources,
2. use bilingual hand-made dictionaries,
3. use Italian as intermediate language,
4. use translation units at sentence boundaries,
5. use different tagsets for source language (SL) and target language (TL).

The first task we completed was that of collecting as much text as possible from the web and from people collaborating on a voluntary basis. Texts collected were then homogenized as to the orthography. Obviously, texts belonging to different varieties were kept separate. As a whole, we collected texts for 200,000 tokens. This was then used to compile frequency lists. The lists were then the basis for the wordform lexicon of Veneto which we compiled following similar lexica we have available in our laboratory, for Italian, English, German and French. The wordform lexicon has been compiled on the basis of the one of Italian, thus comprising in each entry the corresponding Italian wordform and lemma. Semantic and syntactic properties of the Veneto wordform would then be derived directly from the Italian fully specified subcategorized lexicon.

We then normalized a big – 50,000 entries - translation lexicon containing lemmas of Veneto paired with Italian and English. This lexicon will then be used to generate all wordforms of Veneto in this year activities; it is also our current task, the implementation of a morphological analyser for Veneto. The need of the analyser is clear if we think that Veneto makes use of enclitics as Italian and other Romance languages do – more on this topic below.

Eventually, as noted above, we used parallel English-Italian texts to derive multiwords that could then be matched with those present in the Veneto-English parallel texts. We worked at the creation of a small corpus of parallel Veneto-English texts translated by people collaborating at the project. The number of occurrences reaches 40,000 entries but the topics treated in the texts are not homogeneous: children stories and American history. From these materials we managed to collect a small dictionary of 200 multiwords which include very frequent function multiwords, like adverbial and prepositional locutions.

3.1. Using pre-existing tools

As mentioned above, we intend to produce a Transfer based translation which takes advantage of Italian structure similarity to Veneto in order to generate syntactic structural representation and a logical form, to be used in the transfer module. To this aim we intend to use our parser of English that will produce structural representations of some parallel Italian-English corpora – possibly the Europarl corpus. The syntactic structure will then be used to produce two parallel treebanks that together with the word-level alignment should allow us to derive useful information as to the structural correspondances between translations.

At runtime we intend to produce structural syntactic representations, which are basically at constituency level including head modifiers. Best order of the English translation will be at first derived from the Google's English terabyte ngram corpus where we only kept occurrences higher than 1 and came up with some 600,000 entries. We use ngrams also to choose among ambiguous translations. This approach is close to [4] and does not need the generation of a language model. After local consistency has been checked we may need to displace constituents according to the transfer model.

At a preliminary level with tested our prototype with questions in Veneto which show a high level of isomorphism with English. We are developing a scoring function that will allow us to take decisions on how to use context in order to choose the best translation for highly ambiguous cases. Best scores will be reserved to multiwords – more on this below.

4. Symbolic vs. Statistical processing

Problems related to Veneto translation into English and viceversa are very close to those encountered when translating from/into Italian. Basically we can think of the following most interesting types of problems:

a. Subject Clitic Doubling
b. Complementizer Doubling in questions
c. Amalgams (prepositions + article; verb + enclitic)
d. Order of clitics dative/accusative
e. Ambiguous 3rd person singular/plural inflection in present tense
f. Proper Noun preceded by article
g. Subject clitic erased with unambiguous verb inflection (1st sing/plur)
h. Subject adjoined as enclitics in interrogative sentences

Let's look at some examples taken from the website of Dialect Syntax (http://asis-cnr.unipd.it/):

(1) Go poduo dargheo. /I managed to give it to him/her
(2) Ti te parli massa. /You speak too much
(3) No so cossa che fassa e£ Giani /I don't know what John is doing
(4) Partito de boto? /Do you leave at once?
(5) I bocia i magna £e carame£e. /The kids eat sweets
(6) Qua ghe dorme e£ Giani. /Here sleeps John
(7) Dime chi che xe vegnuo. /Tell me who has come
(8) Cossa xe che i fa? /What do they do

What we get here is the lexically unexpressed subject pronoun; then we have a dative clitic pronoun "GHE" which is ambiguous between feminine and masculine. This clitic must be detached from the verb and separated from the accusative "o" or "lo"/it. Most importantly, the order of "accusative/dative" case which is required in English sequences of pronouns, in Veneto is reverted and is identical to Italian.
Another case of ambiguity which requires additional information is constituted by 3rd person plural and singular verb form which are identical. Now we know that English only remarkable morphological marker is the "S" for the singular third person of present tense. In this case, the agreement needs to be recovered from the subject if linguistically expressed, or else from the context. One such case is presented below.
The presence of these features in Veneto do not guarantee the effectivity of statistical models due to the high sparcity of data. Here below we present the structure of our system for syntactic analysis which is used to produce a syntactic representation of English - and Italian if needed.

4.1. An A-As Hybrid Parser

Our parser has been presented in detail lately in a number of papers [5,6] and has achieved 90% recall on the Greval Corpus and 89% recall on the XEROX-700 corpus, this latter test limited only to SUBJ/OBJ GRs – f-score 78%. As in most robust parsers, we use a sequence or cascade of transducers: however, in our approach, since we intend to recover sentence level structure, the process goes from partial parses to full sentence parses. Sentence and then clause level parsing are crucial to the right assignment of Arguments and Adjuncts (hence A-As) to a governing predicate head. This is paramount in our scheme which aims at recovering predicate-argument structures, besides performing a compositional semantic translation of each semantically headed constituent.
The system is organized into twelve layers as described below:

- Tokenizer produces input sentence which is a list of tokens obtained from the input text by sentence splitting;
- Tagger associates lexical categories to words from dictionary lookup or from morphological analysis;
- Tag disambiguation with finite-state automata and the aid of lexical information;
- Head-based Chunk building phase;
- Recursive argument/adjunct (A/A) constituent building procedure as a list of syntactic-semantic structures with tentative GFs labels, interspersed with punctuation marks;
- Clause builder that takes as input the A/A vector and tries to split it into separate clauses;
- Recursive clause-level interpretation procedure, that filters displaced or discontinuous constituents;
- Complex sentence organizer which outputs DAG structures;
- Logical Form with syntactic indices and Semantic Roles;
- Transducer from DAGs to AHDSs by recursive calls;
- Pronominal Binding at clause level followed by Anaphora Resolution at intersentential level;

- Semantic Module to build propositional level feature vectors, which also contain discourse relations.

We would like to define our parser "mildly bottom-up" because the structure building process cycles on a subroutine that collects constituents until it decides that what it has parsed might be analysed as Argument or Adjunct. This proceeds until a finite verb is reached and the parse is continued with the additional help of Verb Guidance by subcategorization information. Punctuation marks are also collected during the process and are used to organize the list of arguments and adjuncts into tentative clauses.

The clause builder looks for two elements in the input list: the presence of the verb-complex and punctuation marks, starting from the idea that clauses must contain a finite verb complex: dangling constituents will be adjoined to their left adjacent clause, by the clause interpreter after failure while trying to interpret each clause separately. The clause-level interpretation procedure interprets clauses on the basis of lexical properties of the governing verb: verbless clauses or fragments are dealt with by adding a default BE dummy predicate.

The final processor takes as input fully interpreted clauses which may be coordinate, subordinate, or main clauses. These are adjoined together according to their respective position. Care is taken to account for Reported Speech complex sentences which require the Parenthetical Clause to become Main governing clause. Specialized procedures are used to deal with non-declarative non-canonical structures like Questions, Imperatives, sentences with Reported Direct speech, Clausal Subject sentences and extraposed That-clause fronted sentences. Fragments are computed at the end as a default strategy.

4.2. Parsing and Robust Techniques

As far as parsing is concerned, we purport the view that the implementation of a sound parsing algorithm must go hand in hand with sound grammar construction. Extra grammaticalities can be better coped with within a solid linguistic framework rather than without it. Our parser is a rule-based deterministic parser in the sense that it uses lookahead to reduce backtracking. It also implements Finite State Automata in the task of tag disambiguation, and produces multiwords whenever lexical information allows it. In our parser we use a number of parsing strategies and graceful recovery procedures which follow a strictly parameterized approach to their definition and implementation. Recovery procedures are also used to cope with elliptical structures and uncommon orthographic and punctuation patterns.

The grammar is equipped with a lexicon containing a list of fully specified inflected word forms where each entry is followed by its lemma and a list of morphological features, organized in the form of attribute-value pairs. However, morphological analysis for English has also been implemented and used for OutOfVocabulary words. The system uses a core fully specified lexicon, which contains approximately 10,000 most frequent entries of English, where every predicate – be it verb, noun, or adjective – is annotated for Syntactic Category, Aspectual Category, Semantic Category; then the list of subcategorized arguments follows (if any exist), each argument being specified by Syntactic Constituency, Grammatical Function, Semantic Role and a list of Semantic Features from a set of 75, the same that we used to relabel WordNet . In addition to that, there are all lexical forms provided by a fully revised version of COMLEX. In order to take into account phrasal and adverbial verbal compound forms, we also use lexical entries made available by UPenn and TAG encoding. Their grammatical verbal syntactic codes have then been adapted to our formalism and is used to generate an approximate subcategorization scheme with an approximate aspectual and semantic class associated to it. Semantic inherent features for OOV words, be they nouns, verbs, adjectives or adverbs, are provided by a fully revised version of WordNet – 270,000 lexical entries - in which we used 75 semantic classes similar to those provided by CoreLex. These are all consulted at runtime. We use these features to induce semantic similarity for two entities whenever at least 2 identical features are matched in their feature list.

Another important element of analysis is constituted by Semantic Roles: we have reformatted all publicly available inventories, such as FrameNet, VerbNet and PropBank, and use them in that order, seen that FrameNet has more specific labels than the other two lexica. However, we also produced our own fully specified lexicon which is accessed before VerbNet.

Our training corpus for the complete system is made up 200,000 words and is organized by a number of texts taken from different genres, portions of the UPenn WSJ corpus, test-suits for grammatical relations, narrative texts, and sentences taken from COMLEX manual.

4.3. Pronominal Binding and Anaphora Resolution

The problem posed by ambiguous unexpressed subject pronouns requires a full-fledged system for anaphora

resolution. One such system is shown in Fig. 1 below, where we highlight the architecture and main processes undergoing at the anaphora level. First of all, the subdivision of the system into two levels: Clause level – intrasentential pronominal phenomena – where all pronominal expressions contained in modifiers, adjuncts or complement clauses receive their antecedent locally. Possessive pronouns, pronouns contained in relative clauses and complement clauses choose preferentially their antecedents from list of higher level referring expressions. Not so for those pronouns contained in matrix clauses. In particular the ones in subject position are to be coreferred in the discourse. This requires the system to be equipped with a History List of all referring expressions to be used when needed.
It is just this mechanism that will allow the system to find appropriate antecedents for unexpressed subject pronouns which will automatically instantiate features like number and gender.

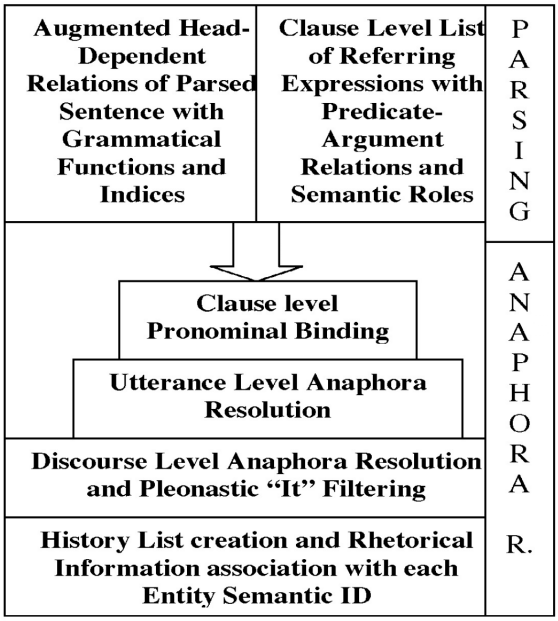

Fig. 1 Anaphoric Processes in GETARUNS

In the system, three levels are indicated: Clause level, i.e. simple sentences; Utterance level, i.e. complex sentences; Discourse level, i.e. intersententially. Our system computes semantic structures in a sentence by sentence fashion and any information useful to carry out anaphoric processes needs to be made available to the following portion of text, and eventually to the Semantic Evaluation that computes entailment. We will comment a number of significant examples to clarify the way in which our system operates.

5. A walkthrough example

We will now show an example and follow it step by step. The sentence is taken from a publicly available database of examples organized by linguists on Italian dialects. The orthography itself is the one used by a linguist and has a strong Venetian bias.

(2) Qualo xé che te piaxe de pì?

FIRST TRANSLATION STEP:
ITALIAN + FEATURES

```
'cua£o'-[quale-['cua£o'-'cua£o'],
         quale-['cua£o'-3-sing]],
xé-[è-[ésar-ind-pres-3-sing]],
ke-[che-ke]
te-[t-[te-2-sing-[dat]],
         te-[te-2-sing-[acc]],ti-[te-2-sing-[dat]]]
piaxe-[piacciono-[piàxer-ind-pres-3-plur],
         piace-[piàxer-ind-pres-3-sing]]
de_pì-nil
?
```

As can be noticed from the output, this level takes care of orthographic mismatches, listed below:

Qualo --> cua£o
Che --> ke

Another important transformation is the multiword "de_pì" which is realized by matching sequences of words to the lexicon of multiwords by means of finite-state-automata. Multiwords constitute unambiguous lexemes which can be translated directly without any intervening additional step.

Second TRANSLATION STEP: Reordering by Transfer Rules

Here no transfer takes place and no chunks are created because the question does not need reordering to take place. So the weighting is only applied to single words rather than to chunks.
As said above, in example (1), the order of enclitics in Veneto should be reversed to suit English order of complement pronouns. However, also this case would not require the use of chunking, but could be realized by ad hoc procedures. We intend to use chunk information only when it is absolutely indispensable. In particular, the presence of unexpressed subject pronouns can only be treated when chunks are computed and the information related to the lack of subject appears clearly from sentence structure. In

order to solve the problem of number ambiguity for third person pronouns, we shall have to derive the information from previous stretch of text. In other words, the empty subject pronoun shall have to be coreferred to some antecedent which will instantiate the number. This is made possible by the Topic Hierarchy, where the Main Topic will be chosen according to semantic feature selection mechanisms.

Third TRANSLATION STEP: Weighting Single Lexical Lookup

What we do at this step is assigning weights to the output of the lexical lookup phase in order to be able to evaluate what word will require more attention in the final step. Basically, the weights are assigned to Italian equivalents which could then be used to select the best structure to match with the output translation. Every time the system spots ambiguity at category level and/or at morphological level, this will contribute a multiplication effect.

'cua£o'-[quale-['cua£o'-'cua£o']-1,
quale-['cua£o'-3-sing]-2]-3
xé-[è-[ésar-ind-pres-3-sing]-2]-2
ke-[che-ke]-1
te-[t-[te-2-sing-[dat]]-2, te-[te-2-sing-[acc]]-2,
ti-[te-2-sing-[dat]]-2]-16
piaxe-[piacciono-[piàxer-ind-pres-3-plur]-2,
piace-[piàxer-ind-pres-3-sing]-2]-8
de_pi-mw-0,
?

This is why the strings associated with “te” and “piaxe” are weighted higher than their linear sum.

Fourth TRANSLATION STEP: English word-level translation pairs

'which_one,which', is, what, you, like, more, ?

As can be noticed, the translation pairs are all unambiguous expect one, the one associated with “cua£o”. So this is solved by accessing Google’s trigrams in the following translation step.

Fifth TRANSLATION STEP: Choice of best translation pairs using Google's trigrams

which, is, what, you, like, more, ?

6. References

[1] Manlio Cortelazzo: "Noi Veneti - Viaggi nella storia e nella cultura veneta..." Revisione didattica di Daniele Cunial, illustrazioni di Marta Tonin e Charlotte Scimemi - Regione del Veneto, Cierre Edizione, 2001, pp.128

[2] Michael Carl, et al. 2008. METIS-II: low resource machine translation, in Machine Translation, 22:67–99.

[3] Mamoru Komachi Yuji Matsumoto, Masaaki Nagata, 2006. Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure, in Proceedings IWSLT 2006, 77-82.

[4] Vamshi Ambati, Alon Lavie, 2007. Occurrence Based Statistics in Machine Translation, 11-731 Spring 2007, Project Report.

[5] Delmonte R., 2007. *Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*, Nova Science Publishers, New York.

[6] Delmonte R., 2009. *Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution*, Nova Science Publishers, New York.