Conclusion

This article proposes a computational aspectual verb classification for Estonian ınsitive change of state verbs. This classification accommodates the systematic ımpatibility of verb classes with certain clausal aspectual object case marking ıtterns. I apply the Lexical Functional Grammar (LFG) methodology. Clausal pect is understood in terms of boundedness. A clause or a sentence is bounded if describes an event with clear boundaries. Clausal boundedness is encoded in the rm of features at the LFG's syntactic level of f(unctional)-structures. This article ıdies those aspect-related attributes and values that transitive change of state rbs contribute to the f-structure. The lexical entries for transitive verbs are ɔvided with specified or underspecified boundedness features in the proposed 'G lexicon.

In this framework, if a verb is called bounded, then its functional specifications ıtain a boundedness feature. This means that these verbs are not boundable any ıre and the range of aspectual case marking possibilities is restricted. If verbs are ındable, their boundedness feature is underspecified. As clausal aspect is deled in terms of the unification of boundedness features in the f-structure, the ısibility of the unification of features with different values explains the wider ge of case marking patterns. In my model, the features become fully specified in process of the unification with the features of case-marked objects.

Verbs fall into aspectual classes, distinguished from each other according to pattern of the attributes and values in the functional specifications of the verbs' ical entries. This verb classification is suitable for accounting for the interaction ween Estonian aspect, verbs, and case.

'erenes

dlinger, Rachel and Louisa Sadler. 2004. 'Tense Beyond the Verb: Encoding Clausal Tense/Aspect/Mood on Nominal Dependents.' *Natural Language and Linguistic Theory* 22. 597–641.

ım, Anne. 2004. *Relations between Estonian verbs, aspect, and case.* PhD thesis, Budapest.

kuyl, H. 1993. *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure.* Cambridge: Cambridge University Press.

Computational aspects of an automatic recognizer of Italian clitics

Marco Tomatis
Università degli Studi di Torino
Via Sant'Ottavio 20 10124 Torino
m-tomatis@tiscali.it

Introduction

The aim of this paper is to present the main features of a computational system for the electronic recognition of Italian clitics.

One of the many problems which may be encountered when preparing a corpus[1] for further (automatic or manual) analysis lies undoubtedly in the so-called text tokenization; the splitting of different words (or lexical units) from all those non-alphabetic graphic signs they may be tied to.[2]

When handling morphologically rich languages like Italian, the problem can not simply be solved by using a text pre-processor: the solution may require considering a more specific level, that is, the morphological structure of the very word. Indeed, for a proper interpretation of the linguistic data of a corpus, the researcher is often obliged to extend the tokenization process even within different words in order to isolate the very word from the clitic it is tied to.

Such a process is, in most cases, very difficult to manage; moreover, when handling large corpora, it has to be done automatically as far as possible. So, to make these tasks easier, an automatic clitic recognising program has been developed.

The system, wholly implemented using a procedural scripting language named "GAWK"[3], acts on an already tokenized text. For it to work correctly, it requires a very complete list of flexed Italian words; obviously without clitics. Since the system is basically founded on linguistic rules, the existence of such a list of words provides a rapid way to check the linguistic hypotheses that are inferred by the rules themselves.

---

[1] Barnbrook (1996); Kennedy (1998).

[2] Grefenstette (1999)

[3] For the complete guide, please read the official Gawk manual:
GAWK: Effective AWK Programming: A User's Guide for GNU Awk. 3rd edition.
Free Software Foundation, Inc. 2001.
The whole manual can be freely downloaded from the address:
http://www.gnu.org/software/gawk/manual/gawk.html
It is also available online at the address: http://it.tldp.org/man/man1/awk.1.html

Methodological approach

The methodological approach, which has been adopted to develop the system, is in some ways innovative. Studies conducted to date on the clitics phenomenon focused attention largely on the leading element, the verb; consequently the clitics issue has been examined mainly from a syntactic or semantic point of view (Borer, 1986) or, in some cases, by adopting a lexical framework which tried to reconduct the behaviour of clitics to that of suffixes (Monachesi, 1999). The approach discussed in this paper, instead, has moved the focus towards the enclitic piece of word only, trying as far as possible to track down the main features of Italian clitics according to their ability to select the verb tense they are attached to. The study of Italian clitics form and behaviour led to drawing up a resumptive table of their main features. For greater terminological clarity, in this paper the general term "clitic" is referred both to simple particles and to multiple clitics chains structured starting from simpler bits, unless stated otherwise.

Clitics	Conjugation	Tense	Number of letters	Final Letter
ccela	1 - 3	imperative	2	A - I
ccele	=	=	=	=
cceli	=	=	=	=
ccelo	=	=	=	=
ccene	=	=	=	=
mmela	=	=	=	=
mmele	=	=	=	=
mmeli	=	=	=	=
mmelo	=	=	=	=
mmene	=	=	=	=
mmici	=	=	=	=
mmiti	=	=	=	=
ttela	=	=	=	=
ttele	=	=	=	=
tteli	=	=	=	=
ttelo	=	=	=	=
ttene	=	=	=	=
cci	=	=	=	=
lla	=	=	=	=

lle	=	=	=	=
lli	=	=	=	=
llo	=	=	=	=
mmi	=	=	=	=
mme	=	=	=	=
tti	=	=	=	=
gli	1 - 2 - 3	infinitive gerund imperative past participle	> = 2	A - E - I - O - R
gliela	=	=	=	=
gliele	=	=	=	=
glieli	=	=	=	=
glielo	=	=	=	=
gliene	=	=	=	=
ci	=	=	> 2	=
cela	=	=	=	=
cele	=	=	=	=
celi	=	=	=	=
celo	=	=	=	=
cene	=	=	=	=
mi	=	=	=	=
mela	=	=	=	=
mele	=	=	=	=
meli	=	=	=	=
melo	=	=	=	=
mene	=	=	=	=
ti	=	=	=	=
tela	=	=	=	=
tele	=	=	=	=
teli	=	=	=	=
telo	=	=	=	=
tene	=	=	=	=
vi	=	=	=	=
vela	=	=	=	=
vele	=	=	=	=
veli	=	=	=	=
velo	=	=	=	=
vene	=	=	=	=
mici	=	=	=	=

tici	=	=	=	=
glici	=	=	=	=
leci	=	=	=	=
vici	=	=	=	=
la	=	=	=	=
le	=	=	=	=
li	=	=	=	=
lo	=	=	=	=
ne	=	=	=	=
si	=	indicative infinitive gerund present part. past participle subjunctive	=	A - E - I - O - N - R
sela	=	infinitive gerund past participle	=	A - E - I - O - R
sele	=	=	=	=
seli	=	=	=	=
selo	=	=	=	=
sene	=	=	=	=
misi	=	=	=	=
tisi	=	=	=	=
glisi	=	=	=	=
lesi	=	=	=	=
cisi	=	=	=	=
visi	=	=	=	=

Table 1

The table above is divided into five fields. The first column on the left includes the list of all the standard modern Italian clitics. The second and the third fields include respectively the conjugation and the tense of the verb which the clitic can be attached to. The fourth field, instead, lists the syllabic structure of the leading verb by specifying the exact number or the minimum threshold of letters which it has to be made of in order to be selected by a particular set of clitics. Finally, the fifth field lists the different verbal flexions which the clitic can be tied to.

Analysis of data

The table shows that Italian clitics can be divided into two main groups; those starting with a geminate consonant and all the rest, which may be further classified following their own intrinsic features.
Although it is quite easy to spot the rules useful to handle the set of clitics that start with a geminate consonant, for the bigger group made of simple and articulate clitics a more thorough examination of their verb selection behaviour is required.
After a detailed analysis of the data displayed in table 1, it is possible to group the clitics into homogeneous macro areas. A first area should include the clitic "gli" and its articulate forms "gliela", "gliele", "glielo", "glieli" and "gliene". Such clitics show a very interesting behaviour; like a hybrid entity they share the features of both the clitics having a geminate consonant (which take monosyllabic imperatives) and a great part of the remaining. Yet two forms which do not belong to such set exist; they are "glici" and "glisi". In fact the latter behave differently from the other compound clitics starting with "gli"; in particular they cannot be tied to monosyllabic verbs. For this reason they should be included into different areas.
Another large group includes the simple clitics "ci" "mi" "ti" "vi" and their articulate forms which take the pronouns "lo" "la" "li" "le" and the adverb "ne". Such set of clitics is different from the previous one because it can not take monosyllabic imperatives but only plurisyllabic ones, plus verbs conjugated in the infinitive, gerund and past participle forms. This group may also include the subset filled with those compound clitics starting with a personal pronoun "mi" "ti" "gli" "le" "vi" and ending with the second adverbial element "ci", though such particular combinations appear more and more rarely in contemporary Italian (i.e. "porta***mici***").
Another area includes the simple form "si" only. This clitic proves to have a different selectivity level compared to the other ones because it takes a wider range of verb conjugations. As a matter of fact it can be tied to the infinitive, gerund, present and past participle and the third person singular and plural of the simple present tense (e.g. "compra***si***", "vendon***si***", etc.), plus the third person singular of the present subjunctive tense (e.g. "legga***si***", etc.).
Finally, the last clitics that may be grouped in a homogeneous set are the compound forms of the personal pronoun "si"; they can only take verbs conjugated in the infinitive, gerund and past participle forms.
The kind of approach described so far has helped to formulate a set of handy rules to distinguish, within the specific lexical entry, what is to be considered a real clitic from what is simply a bare segment of the whole word (e.g. ved***ine*** - "see of it" vs. ped***ine*** - "pawns"). A brief description of the general setting of the system and its recognition rules follow.

Software architecture

The automatic Italian enclitic tracking system "ClitRec" examines the longest clitic blocks first. When the whole set of rules returns a positive result, the clitic particle is divided from the word itself and marked by a univocal sign, then the result is printed out and the control routine moves to the next word in the text line. Otherwise what may happen is that after ending the entire analysis procedure, the system cannot match the hypothetical clitic with the list of possible Italian clitics. So the whole word is printed with no changes and the routine starts again from the next lexical unit. Inversely, once the software successfully matches the probable clitic with one in the list, the linguistic validity of the clitic is further evaluated against the set of linguistic rules. After that, the software may print out the positive result or it may carry on its task until it finishes all the linguistic material to be examined.

Pre-processing activities

In order to enhance the computational efficiency of the system discussed in this paper, as well as optimizing the algorithm used in the clitic recognition activity, a number of pre-processing rules have been defined. Since their main task is to filter out irrelevant or noise producing lexical material, the said rules accomplish the following actions:

- Filtering of all those words whose final part does not match with a possible clitic.
- Filtering of those words containing clitic elements which are ambiguous due to the lack of graphic stressing marks (e.g. "mangia***telo***" vs. "mangiate***lo***", "guarda***tene***" vs. "guardate***ne***")
- Filtering of those forms identical to compound prepositions (e.g. "da***llo***", "da***gli***", etc.)
- Checking of the existence in the Italian lexicon of the piece of word obtained by depriving the whole word of the hypothetical clitic element. The word would be excluded from further analysis if this check returned a negative result. This check is not run both on the verbal forms conjugated in the infinitive form and in the third apocopate plural person of all the tenses involved.
- Exclusion of those words containing probable, but not real, clitics.
    - In the case of a word ending with the vowel "a" (e.g. "peda***la***", "affi***la***", etc.) the checking system provides for the addition to the end of the word of the part corresponding to the infinitive verbal flexion, followed by a check of the existence of such new form into the Italian lexicon ("peda**lare**", "affi**lare**", etc.)
    - Cases different from the one above require further tests to ascertain the existence of the whole word in the lexicon. The system will exclude the existence of a clitic in the word if, after substituting its final vowel with those bearing a particular value of gender and number, the new form proves to be part of the Italian lexicon used by the software. (e.g. "fifo***ne***" - "fifo**na**" "fifo**ni**"; "feri***ti***" - "feri**ta**" "feri**te**" "feri**to**")

Analysis rules

After having described the general structure of the software and the first stage filter rules, a sample description of some linguistic rules used by the program follow.

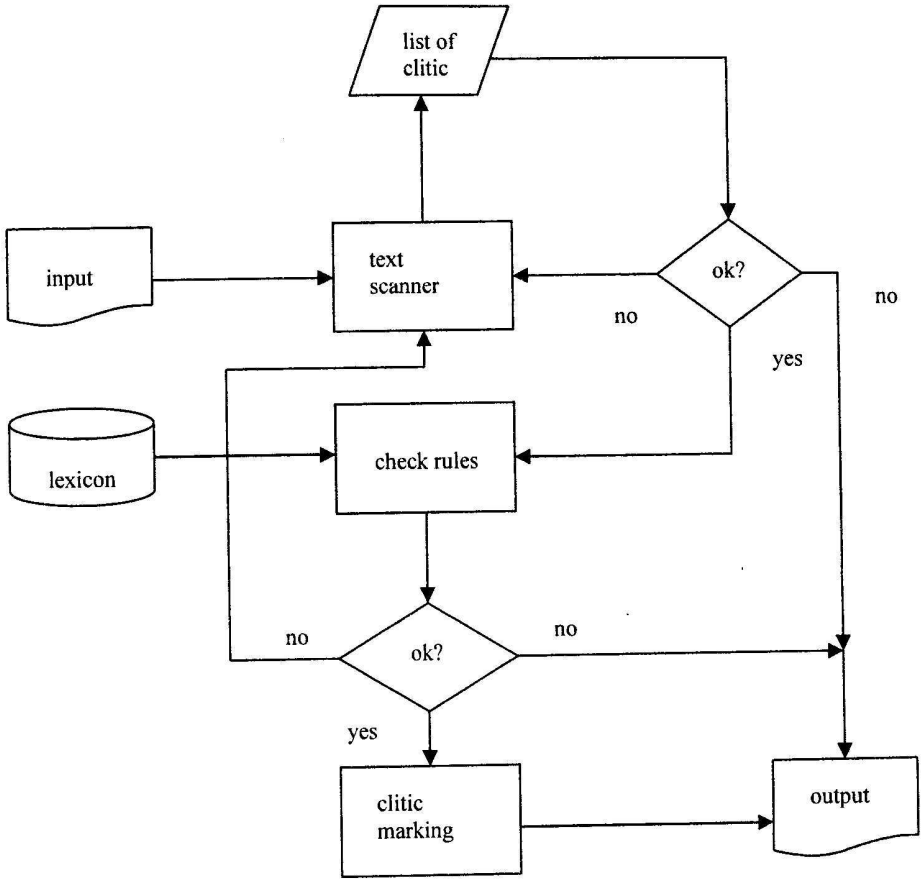
- Clitics starting with geminate consonant: they can only take monosyllabic imperative verbs or their iterative forms (e.g. "di***mmi***", "*ri*di***mmi***", etc.). This rule avoids automatically analysing those lexical forms which prove ambiguous due to their transcategorization (i.e. "fa***lla***", "fa***llo***", etc.)[4]
- Clitic "**gli**", simple form. It is always recognised as a clitic when found tied to the monosyllabic imperatives "di", "da", "fa" and their own iterative forms. The imperative conjugation of the verb "andare" ("va") has not been taken into account by this rule in order to avoid ambiguity at morphologic level between that word and the present subjunctive singular of the verb "vagliare" ("vagli")
- Clitic "**gli**", compound forms. They are always recognised as clitics when found tied to the monosyllabic imperatives "di", "da", "fa", "va" and their own iterative forms.
- Clitic "**ne**": rule to clear the ambiguity with augmentative and diminutive forms. This check acts substituting the last vowel of the piece of word in exam, previously deprived of the potential clitic, with the 2nd and 3rd conjugation of the gerund flexion ("endo"). For clarity's sake let's examine the following words as examples: "prendi***ne***" and "costi***ne***". Adopting the criterion explained before, the particle "ne" belonging to "costine" will be never analysed by the system as a clitic; instead it will the "ne" part of "prendine" simply because within the Italian reference lexicon this particular rule will find the word "prend**endo**", while something like "cost**endo**" will never be found. Moreover, it is important

[4] The term "transcategorization" means that a particular word can belong to different part of speech (or grammar categories). In the example given, the term "falla" belongs to both categories of nouns and verbs.

to notice that, in order to avoid misleading analysis due to non-existing words that may be generated by spelling errors (i.e. "dist***ine***" instead of "distinte"), the program does not limit itself to the tests on the lexicon discussed before, but it runs a specific one which substitutes the gerund flexion with the 2nd and 3rd conjugation of the simple imperfect tense flexion. Even though this further check may seem redundant, it is in fact extremely significant and useful. Although the last example takes into account a badly formed Italian word, the part of word that is deprived from the potential clitic particle ("disti") is exactly the same as the 1st, 2nd and 3rd persons singular of the present subjunctive of the verb "distare". Since the lexicon is not supposed to be a POS (part of speech) tagged text, the system could not but recognize the particle "ne" as a real clitic, even though such kind of clitics can not tie themselves to verbs conjugated in the subjunctive. So, given these premises, if the rule simply substituted the final vowel "i" with the whole gerund flexion "endo", the newly formed word "dist**endo**" would correspond to the 1st person singular simple present conjugation of the verb "distendere", which would again lead to the wrong result of validating the hypothesis that the particle "ne" is a real clitic and not a mere part of the word. As a matter of fact, only the test which substitutes the gerund flexion with the 2nd and 3rd persons singular of imperfect (*"dist**iva**"; *"dist**eva**") and then use the Italian lexicon to test their existence allows the system to avoid such a mischievous error. Finally, it is important to remark that the test uses both the flexion of 2nd and 3rd conjugation of the simple imperfect tense because it is not possible to infer the correct conjugation a verb belongs to simply by examining its imperative form (e.g. "bev***ine***" "bev**endo**" "bev**eva**" - "apr***ine***" "apr**endo**" "apr**iva**").

- Rules for clitics which are tied to morphologically irregular verbs in the imperative and subjunctive tenses[5] (i.e. "sappi" - "sappia", "siedi" - "segga", etc.). Since in the Italian lexicon such verbs are very few in number, if only the piece of word deprived of the potential clitic should be found in the reference lexicon, but not the complete word, it is possible to infer with a good error margin that the specific particle under exam is a real clitic, not merely the final part of a word.

[5] In Italian, when moving from the subjunctive to imperative tense, regular verbs simply change their final vowel only. The 1st person conjugation provides for the verb to take a flexion "a" in the imperative form (i.e. "guard**a**") and a flexion "i" in the subjunctive tense (i.e. "guard**i**"). On the contrary, the 2nd and 3rd conjugations behave in the opposite way; verbs take a "i" in the imperative tense (i.e. "prend**i**") and an "a" in the subjunctive tense (i.e. "prend**a**"). All those verbs which do not behave in the standard way must be managed using specific rules.



Flow chart of the clitics recognizing system

Future developments

The system described in this paper is still in working progress; its quality could be improved by using a preliminary stochastic part of speech tagger in order to disambiguate words which could not be treated in other ways (i.e. "mangia***tene***" - "eat some of it" (2nd person singular) vs. "mangiate***ne***" - "eat some of it" (2nd person plural). The said tagger should work together with a morphological analysis

system, which is currently being developed, to help recognize and divide possible prefixes from the lexical stem (i.e. "**stra**dilungarsi" - "to linger unduly") in order to free the clitic recognizing system from the need to use a control lexicon of all the words existing in the Italian language.

Conclusion

This paper has described a system which helps to track the pronominal and adverbial enclitic part of word in a non POS tagged corpus. Even though this system is not based on stochastic inference functions, a complex set of rules enables us to reach a rather high analytical level. Finally, the adoption of a comprehensive Italian lexical database helps the different rules to optimise their inference.
The program has been written using a scripting language named "GAWK" which allows the developer to create a fast, portable, easy to maintain program which does not require the installation or running of complex procedures by the final user.

Bibliographic References

Aarts J. and Meijs W. "Theory and practice in corpus linguistics." Amsterdam & Atlanta: Rodopi 1990.

Barnbrook G. "Language and Computers. A Practical Introduction to the Computer Analysis of Language." Edinburgh: Edinburgh University Press 1996.

Borer H. (ed .), The Syntax of Pronominal Clitics, "Syntax and Semantics" 19, New York : Academic Press, 1986

Calabrese, A. "I pronomi clitici." In: L. Renzi (ed.) Grande Grammatica Italiana di Consultazione. Vol.1. Il Mulino, Bologna, 1988

Grefenstette, Gregory. "Tokenization." In van Halteren, chap. 9, pp. 117–133 1999.

Kennedy G. "An Introduction to Corpus Linguistics." Longman: London & New York 1998.

Monachesi, P. "A Lexical Approach to Italian Cliticization." Stanford: CSLI Publ. 1999.

Simpson J. & M. Withgott, "Pronominal clitic clusters and templates" in Borer (ed.) 1986: 149-174

Reverse Lemmatizing of the Dictionary of Middle Dutc

(1885-1929)

Using Pattern Matching

John van der Voort van der Kleij

voortkleij@inl.nl

Institute for Dutch Lexicology

Leiden, The Netherlands

Abstract

The Integrated Language Database (ILD), a project of the Institute for Dutch Lexicology (I http://www.inl.nl), will contain various kinds of Dutch language data from the earliest to the most re periods, such as electronic dictionaries, text files and files with linguistic data (lexica). There will be a v balanced selection of sources, linguistic annotation of the texts and the linking of sources. This proje database will be a research tool for various aspects of the Dutch language and culture throughout centuries.
The Dictionary of Middle Dutch (MNW), now available in electronic form, is a classic lexicographic so of nine large volumes that will be incorporated in this database. To link the ca. 74000 entries of dictionary with the corresponding entries of other lexicographic sources in the database (for example dictionary of the Dutch Language on historical principles, WNT) modern Dutch entry forms are b added.
Our paper concerns the links between Middle Dutch wordforms (tokens) in their context with the entri the dictionary. For lemmatizing Middle Dutch texts we need a lexicon covering the paradigms of entries. To build such a lexicon we developed a sophisticated pattern matching program that links the e forms with their paradigmatic types in the quotations. Basis for the matching are the dictionary entries their listed variants.
A wider perspective is, that this lexicon of paradigmatic forms may be extended into a morphole computer lexicon. Of course part of speech information needs to be added. Other Middle Dutch diction will also be exploited, like the Dictionary of Early Middle Dutch (VMNW) and the concise one vol dictionary of Middle Dutch, an excerpt of the MNW and for many articles a revision of its source.