Marco Battista & Vito Pirrelli, ILC-CNR, Pisa

Descriptive Economy and the Morphological Lexicon. On the Use of Interparadigmatic Schemata in a Computational Lexicon for the Morphology of Italian[1]

Abstract

The paper illustrates the general architecture of a computational lexicon for the Morphology of Italian which accomodates the entire range of regularities, subregularities and sheer exceptions of the Italian inflectional system in a concise and elegant way. The tenet of the lexicon architecture is the novel notion of "interparadigmatic schema" and rests on the straightforward observational fact that, beyond the paradigmatic level of organization of the inflectional system of a highly inflecting language, there is a further level of interparadigmatic redundancy which can conveniently be captured and exploited for purposes of lexicon compilation. Our experience shows that a lexicon based on the use of such interparadigmatic schemata compares favourably with other computational lexica where these schemata are ignored.

1. Paradigm-Morphology and Italian inflection

Inflection in Italian grudgingly lends itself to the notion that Morphology is an inventory of unrelated items, called "morphemes", consisting of an arbitrary association of a form (a continuous phonological sequence or string of characters) and a well-defined portion of morphological meaning (either morphosyntactic or lexical). This picture is purported to work well for the description of so-called agglutinating languages such as Turkish but raises several difficulties when applied to inflecting languages like Italian.

Paradigms add a further dimension to morphological analysis: no matter what formal means are put to use to realize a given morphological meaning, paradigms lay stress on the way morphological meaning is systematically structured in language. For example, the Italian verb conjugation system involves the following morphological dimensions: tense, mood, person and number, conjugational class of the verb in question. Morphological formatives vary depending on how these dimensions are combined (from less to more inclusive): there are several combinations of person and number values for the same values of tense,

mood and conjugational class, several tenses for the same mood and conjugational class, several moods for the same conjugational class. A presupposition relation (Wurzel 1989) holds between the formatives associated with different person-number values of the same tense, mood and conjugational class: if an Italian verb takes *-a* as the ending of third-person singular of present indicative, it will also take, among other formatives, *-iamo*, *-ate*, *-ano* respectively for first-, second- and third-person plural of present indicative. The family group of all formatives which happen to be linked with one another through a presupposition relation is commonly referred to as a "paradigm". Each legal combination of paradigmatic dimensions constitutes a slot. Barred some exceptional cases, a paradigm holds for an entire class of words: for example, first conjugation in Italian is the verb inflection paradigm of all verbs with the infinitive form in *-are*.

Besides the intraparadigmatic redundancy enforced through Wurzel's presupposition relation, inflectionally rich languages exhibit also systematic "interparadigmatic" regularities (Spencer 1991). All irregular verbs of Italian, independently of their conjugation class, present at least one stem allomorph alternating with its base within particular paradigmatic slots: e.g. *andare/vado* 'go/I go', *vengo/vieni* 'I come/you come' etc. In all such cases, a systematic relationship holds between a particular stem alternant and a set of paradigmatic slots: for example, if the first-person singular of a verb in the present tense requires a different alternant from the one used for second-person singular, the former and not the latter is required in the third-person plural of the same tense.

2. Stem alternation and Stem Choice in Italian Conjugation

Stem alternation has been accounted for in the literature in a number of ways: 1) as a co-selection between two independently identified morphological units (i.e. a stem and an inflectional ending) (Jensen 1990), 2) as the result of the application of a unique morphological process to a stem (Hoeksema and Janda 1988), and 3) as a phonologically-governed alteration of the stem when in company of particular endings (Halle and Mohanan 1985). All these solutions miss the interparadigmatic dimension of reciprocal conditioning in Italian inflection. Consider the following examples:

a)

teng	o
tien	i
tien	e
ten	iamo
ten	ete
teng	ono

b)

veng	o
vien	i
vien	e
ven	iamo
ven	ete
veng	ono

c)

dolg	o
duol	i
duol	e
dol	iamo
dol	ete
dolg	ono

d)

colg	o
cogl	i
cogl	e
cogl	iamo
cogl	ete
colg	ono

Tenere 'keep' and *venire* 'come' in a) and b) above, while belonging to two different conjugational paradigms (respectively second and third conjugation), exhibit a strikingly similar pattern of stem alternation.

1)

1s	a
2s	c
3s	c
1p	b
2p	b
3p	a

2)

1s	a
2s	b
3s	b
1p	b
2p	b
3p	a

3)

1s	a
2s	
3s	
1p	b
2p	b
3p	a

In grid 1) above the same stem alternants are replaced by the same index, and inflectional endings are abstracted away. Such an abstract paradigm (hereafter "interparadigmatic schema") applies to the present indicative of both *venire* and *tenere* with no further readjustment. Furthermore, exactly the same interparadigmatic schema holds for the present indicative of the verb *dolere*, independently of the considerable variation in phonological content between its stem alternants and those of *venire* and *tenere*. We observed that comparatively few paradigmatic schemata of this sort, underspecified for phonological content, suffice to generalize over the great variety of conjugational classes in Italian (Pirrelli 1993, Battista and Pirrelli 1995).

There appears to exist a straightforward relationship between the number of stem alternants of a verb and the interparadigmatic schema the verb conforms to. Under the assumption that three stem alternants *a*, *b* and *c* are used for the present indicative, they will distribute according to grid 1) above with very few exceptions. Stem *a* occurs in both first-person singular and third-person plural; stem *c* in second- and third-person singular, and so on and so forth. Similarly, a verb with two stem alternants in the present indicative will rarely depart from the schema exemplified in grid 2) above. Schemata 1) and 2) present a nonempty intersection (schema 3) which represents the considerable amount of combinatoric redundancy carried over from a three stem alternant to a

two stem alternant paradigm. In particular, i) the same stem shows up in first-person singular and third-person plural, ii) *b* has the same distribution in the two schemata; iii) *b* occupies the slots left empty by the missing third stem alternant, so that the constraint that stem alternants in paradigm slots 2s and 3s are assigned the same index is not violated. A quick look at the paradigm of completely regular verbs (i.e. verbs with no stem alternation) confirms this trend: *a* disappears and its slots are taken over by *b*. This supports the claim that *b* is the default alternant and normally occupies the slots which are left empty by missing alternants.

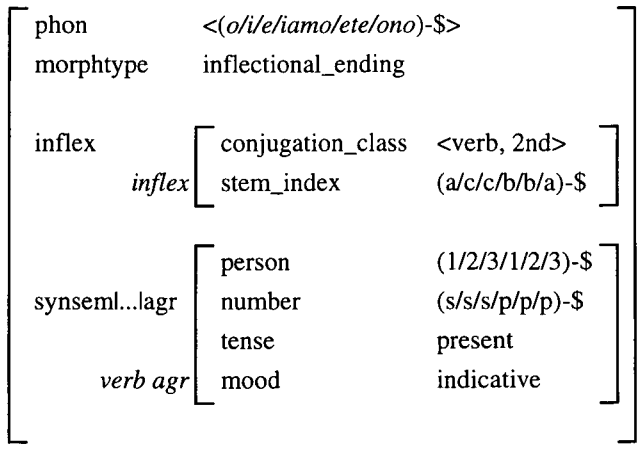
3. Interparadigmatic Schemata and the Lexicon

The descriptive economy one gains in encoding inflectional paradigms in the lexicon through reference to an interparadigmatic schema is considerable. First, the inflectional paradigm for present indicative is represented once in the lexicon through the relevant interparadigmatic schema specified for the maximum number of possibile stem alternants (diagram 4).

4) interparadigmatic schema

a	o
c	i
c	e
b	iamo
b	ete
a	ono

5) second_conjugation_present_indicative



More formally, this information can be encoded through the TFS entry in 5), where the paradigm of present indicative of second conjugation verbs is expressed as a disjunction of inflectional endings (in the attribute "phon"), their indices (in "stem_index") and corresponding paradigmatic slots (expressed as a distributed value in "synseml...lagr"). A feature value can be an atom (e.g. *present*) or a list (between "<>"). Disjunctive

lists are enclosed in parentheses, with items separated by ';' (the simple "or" operator) or '/' in so-called "named disjunctions"[2] (Krieger and Nerbonne 1991). Named disjunctions are used in 5) to enforce the disjunctive covariation between inflectional endings (*o/i/e/iamo/ete/ono*), indices (a/c/c/b/b/a), and person and number values. This means that – say – when the feature "phon" takes the value *o*, "stem_index" gets *a*, "person" gets *1*, and "number" gets *s*.

An example of lexical entry of a verb stem which combines with 5) is given in 6) for *tenere*.

6) *tenere*

phon	<(*ten/teng/tien/tenn/ter*)-$>		
morph-type	stem		
inflex	conjugation_class	<verb, 2nd>	
inflex	stem_index	((b;f)/a/c/d/e) -$	

"Phon" takes as a value the disjunctive list of *tenere* stem alternants. Hereafter we will refer to a disjunctive list of stem alternants as the "alternation pattern" of a verb. 6) says that the stem *ten* covaries with indices *b* and *f* (where *f* is the index for past participle), *teng* with *a*, *tien* with *c* etc.

Following Stump (1995) we call "stem indexing" the covariation between a stem alternant and a paradigm index. Different values of "stem_index" in verb entries, that is different ways of grouping indices separated by '/', correspond to different patterns of stem indexing. 5) combines also with verbs with fewer stem alternants with a minimum of fiddling with the verb "stem_index". First, attested stem alternants are assigned the appropriate index; indices for missing alternants are then associated with the default alternant. This is illustrated in 7) below for the two alternant verb *conoscere* 'know', where the default stem *conosc* is assigned all paradigmatic indices but *d*, independently associated with *conobb* for past tense formation.

7) *conscere*

phon	<(*conosc/conobb*)-$>		
morph-type	stem		
inflex	conjugation_class	<verb, 2nd>	
inflex	stem_index	((*a;b;c;e;f*)/*d*)-$	

A fully inflected word form is yielded as the mother node of a binary branching tree structure whose daughters are one entry of type "verb stem" (such as 6 and 7) and one entry of type "inflectional ending" (such as 5): only entries which are specified for unifiable "inflex" feature structures eventually combine.

There is a surprisingly limited variety of attested interparadigmatic schemata in Italian: although three stem alternants could potentially combine in 3^6 different ways for present indicative, barred very few truly exceptional cases, Italian exhibits only the schema in 4)[3]. Like interparadigmatic schemata, also patterns of stem indexing are fewer than expected. In theory, there are 6•5 = 30 possible different ways of combining the index of a missing alternant with five attested indices. Out of them, only two such ways are found in Italian. All in all, patterns of stem indexing are considerably fewer than patterns of stem alternation. As shown in section 2, five alternant verbs such as *tenere* and *dolere* exhibit the same indexing pattern in spite of the substantial difference in phonological content of their respective stem alternants. The same holds for three alternant verbs such as *friggere* 'fry' and *vincere* 'win' in entries 8) and 9) below, whose "stem_index" values are identical, contrary to their "phon" values.

8) *friggere*

phon	<(*frigg/friss/fritt*)-$>		
morph-type	stem		
inflex	conjugation_class	<verb, 2nd>	
inflex	stem_index	((*a;b;c;e*)/*d*/*f*)-$	

9) *vincere*

phon	<(*vinc/vins/vint*)-$>	
morph-type	stem	
inflex	conjugation_class	<verb, 2nd>
inflex	stem_index	((*a;b;c;e*)/*d/f*) $

10) Macro_IND_3stems_1[]

inflex|stem_index	((*a;b;c;e*)/*d/f*)-$

It makes sense to exploit these regularities for lexicon compilation through use of lexical macros, such as the one in 10), that is functions from arguments to pieces of lexical coding. 10) above serves the purpose of spelling out the appropriate "stem_index" value of a verb with 3 alternants when invoked by the lexicon writer for lexicon compilation.

4. Stem indexing and stem formation

Consider the alternation pattern in 11) which accounts for verb form pairs such as *crocifiggere/crocifissi* ('crucify/ I crucified'), *friggere/frissi* ('fry/ I fried'), *infliggere/inflissi* ('inflict/ I inflicted'):

11) X-*gg* ↔ X-*ss* (for second conjugation verb stems only)

11) is indexed differently depending on the verb stem: "X-*ss*" occurs also in the past participle of *crocifiggere*, *prefiggere*, *infiggere* etc. (*crocifisso*, *prefisso* etc.), but alternates with "X-*tt*" in the past participle of *friggere*, *infliggere*, *correggere* etc. (*fritto*, *inflitto* etc.). This evidence seems to lend support to Greg Stump's hypothesis, known as Indexing Autonomy Hypothesis (IAH), that "stem pairs exhibiting an identical contrast in formation needn't exhibit an identical contrast in indexing". However, some qualifications are in order here. First, although two verbs such as *infiggere* and *infliggere* share the same stem alternation, their sets of alternants are not coextensive: *infiggere* simply lacks an X-*tt* alternant. Note further that 12) and 13) below, under the assumption that both a) and b) are true, are never violated in Italian:

12) **if** a) X-*gg* X-*ss* **and** b) there is no other stem alternant
then *a; b; c;e* = X-*gg* **and** *d; f* = X-*ss*

13) **if** a) X-*gg* X-*ss* X-*tt* **and** b) there is no other stem alternant
then *a; b; c;e* = X-*gg* **and** *d* = X-*ss* **and** *f* = X-*tt*

This is tantamount to stating the following Paradigm Indexing Hypothesis (PIH): "the indexing schema of a verb stem V is predictable if we know: 1) the number of alternants of V, and 2) the sort of alternation they undergo". The macro 14) below is a practical instantiation of PIH, where '&' is the operator of string concatenation. 14) is able to spell out the entry of *friggere* in 8) above when the argument X is assigned the string *fri*. Clearly 14) presupposes the stipulation of Macro_IND_3stems_1[] in 10) above, which is also independently needed by yet other complex macros such as 15) below, which accounts for the alternation patterns of *vincere*, *torcere* and the like.

14) Macro_PHON_IND_3stems_1 [X]

phon	*<(X&gg/X&ss/ X&tt)-$>*
Macro_IND_3stems_1[]	

15) Macro_PHON_IND_3stems_2[X]

phon	*<(X&c/X&s/X&t)- $ >*
Macro_IND_3stems_1[]	

5. Conclusion

It is a well-known fact that paradigms represent descriptively adequate formal tools for dealing with inflectionally rich languages such as Italian (Stump 1991, Dumitrescu 1992, Carstairs 1992, Spencer 1991). In this paper, we provide a further convincing argument in support to this claim: inflectional paradigms are very natural linguistic classes, which appear to convey a considerable amount of redundancy in the inflectional system of Italian with a compartively sparse repertoire of formal means. We show that this redundancy can be captured in the lexicon by means of so-called interparadigmatic schemata. In particular, interparadigmatic schemata are able to accomodate thorny cases of stem alternation in Italian with a strikingly sparse amount of lexical encoding.

Notes

1. All ideas illustrated here are the outcome of a joint effort. For the specific concerns of the Italian Academy only, V. Pirrelli is responsible for sections 1 and 2, M. Battista for sections 3 and 4.
2. Named disjunctions are lists of elements enclosed between brackets and separated by '/': the name is suffixed and separated by a dash ("-$" in the example at hand). The elements of two identically named disjunctions are made covary in their respective order.
3. For a thorough description of the overall interparadigmatic schema of Italian verbs see Battista and Pirrelli 1996.

References

Carstairs-McCarty A., 1992, *Current Morphology*, London: Routledge.

Dumitrescu C., 1992, Paradigmatic Morphology Modeling and Lexicon Design with MORPHO-2, in *Proceedings of Euralex '92*, pp. 203–212.

Halle M., and K.P. Mohanan, 1985, Segmental Phonology of Modern English, *Linguistic Inquiry*, 16.

Hoeksema J. and R.D. Janda, 1988, Implications of Process-Morphology for Categorial Grammar, in Oehrle, Bach, Wheeler (eds.), *Categorial Grammars and Natural Language Structures*, Dordrecht: Reidel Publishing Company.

Jensen, J.T., 1990, *Morphology: Word-Structure in Generative Grammar*, Holland-USA: J. Benjamins.

Pirrelli V., 1993, *Morphology, Analogy and Machine Translation*, Ph.D. Dissertation Salford University U.K.

Spencer A., 1991, *Theoretical Morphology*, Basil Blackwell.

Stump G.T., 1991, A Paradigm-Based Theory of Morphosemantic Mismatches, in *Language*, 67, pp. 675–725.

Wurzel W., 1989, *Inflectional Morphology and Naturalness*, Dordrecht: Reidel.