STEM SPACES AND PREDICTABILITY IN VERBAL INFLECTION

FABIO MONTERMINI, OLIVIER BONAMI

ABSTRACT: This paper outlines and justifies a general analysis of Romance verbal paradigms. In most Romance languages, verbs present a complex inflectional system, where lexemes are divided into several classes and subclasses, inflectional endings are highly redundant across classes for the same cell, and stems present a high degree of internal variation. These facts make an analysis based on a unique invariant underlying form difficult, as it requires the elaboration of a complex system of rules for deriving surface forms. By contrast, the approach described here aims at reaching economy of description not by minimising the amount of stored information, but by proposing strategies for organising the complex phonological representations attached to lexical units. The tools of this organisation are stems, i.e. phonological subcomponents of sets of inflected forms in systematic co-variation; a stem space, specifying the distribution of stems throughout paradigms; and a stem graph, which encapsulates the predictability relations in the stem graph through a minimal set of connections that are maximally reliable.

KEYWORDS: Inflectional morphology, allomorphy, stems, paradigms, Romance languages.

1. INTRODUCTION[*]

One main challenge of present-day studies on morphology is the elaboration of theoretical models that are compatible with current knowledge on morphological processing and the mental lexicon.[1] This paper aims at contributing to the construction of such models by presenting an overview of

[*] We are grateful to Vito Pirrelli for having given the first impulsion to this paper; we are also grateful to the participants to the workshop *Understanding the Nature of the Mental Lexicon*, Pisa, 24-26 November 2011, where we presented a preliminary version of this paper, for their valuables comments; finally, we wish to thank Gilles Boyé and Basilio Calderone for many discussions of the issues presented here.

[1] Notice that this is the morphological instanciation of what Sag & Wasow (2012) call *Performance-compatible competence grammar.*

recent research the conjugation systems of Romance languages. We will propose a reassessment of the developments morphology has undergone in the last decades as these have inpacted descriptions of Romance conjugation, and will discuss some of the notions that have been elaborated in order to structure inflectional paradigms, such as the morphome (Aronoff 1994; Maiden 2005, 2009), the Stem Space (Pirrelli & Battista 2000; Bonami & Boyé 2003, 2007), and the very notion of paradigm.

The paper is organised as follows: in section 2 we give an overview of some models of morphological analysis, focusing in particular on the dichotomy between abstractive and constructive models. A main characteristic of abstractive models is that they treat the various inflected forms of a lexeme as primitives of morphological description, rather than trying to systematically treat them as derived objects through the application of operations to a single underlying form. We also present data from the conjugation systems of various Romance languages that are problematic for a purely constructive approach that treats all the subparts of complex word forms as realising content. Section 3 is devoted to the justification of one particular abstractive approach, which aims at explicating an inflection system not by minimising the amount of stored information but by organising the complex entries that correspond to the phonological representations of lexical units. The tools for doing so are the stem space, a scheme of the distribution of allomorphic stems within a paradigm, and the stem graph, which connects stems according to their degree of interpredictability. Section 4 situates the current paper in the context of other current work on implicative morphology.

2. MODELS OF MORPHOLOGICAL ANALYSIS

2.1 Abstractive vs. constructive models

Morphological models can be classified in several different ways; see Hockett (1954) and Stump (2001, chap. 1) for two very popular attempts. A particularly insightful distinction is the one made by Blevins (2006) between abstractive and constructive models. In abstractive models, the recurrent elements that are isolated in morphological analysis are the result of abstractions made over full forms, while in to constructive models surface full forms are the results of the combination of basic elements, such as roots and affixes (Blevins 2006: 533). As Blevins points out, it is natural for abstractive models to be word-based, i.e. to take full inflected word forms as primitives. Constructive models are typically morpheme-based, and consider that roots and exponents (most commonly affixes) are similar in nature.

However, even non-morpheme-based models such as *A-morphous morphology* (Anderson 1992), *Paradigm Function Morphology* (Stump 2001) or *Network Morphology* (Brown & Hippisley, 2012) are constructive: each lexeme is associated with one basic phonological representation, a (a root or basic stem) from which full forms are deduced through the application of various operations; lexical listing of extra information (stem alternants or full forms) is a deviation from expectations on the normal business of morphology.

In turn, abstractive and constructive approaches are more readily compatible, respectively, with an enriched or an impoverished view of the mental lexicon. Since what is essential, for abstractive approaches, are the relations between attested full forms, there is in theory no limit to the number of forms that can be stored for each lexical entry, while constructive approaches tend towards a model in which few basic elements are combined through a rich computational system, and thus have a one unit – one form correspondence as an ideal (cf. Derwing & Skousen 1988; Stemberger & MacWhinney 1988; Jackendoff 2002, among others, for an overview of different views of the lexicon).

2.1 Problematic data

According to constructive approaches, then, all the variation actually observed in inflectional paradigms can be reduced to a unit by postulating a (possibly abstract) basic form, from which all other forms can be derived by means of more or less general rules. Table 1 presents an example of such variation, taken from the present indicative of two French verbs showing a high degree of allomorphy.

Person					
1SG	2SG	3SG	1PL	2PL	3PL
tjɛ̃	tjɛ̃	tjɛ̃	tənɔ̃	təne	tjɛn
ʁezu	ʁezu	ʁezu	ʁezɔlvɔ̃	ʁezɔlve	ʁezɔlv

TABLE 1. PRESENT INDICATIVE OF THE FRENCH VERBS TENIR ('KEEP') AND RÉSOUDRE ('RESOLVE')

Of course, it is in theory possible to establish phonological mechanisms suitable to account for the variation observed in the data in Table 1, for instance by postulating, for TENIR, a rule of diphthongation / fronting in a specific position, or by postulating the existence of non-realised elements producing a denasalisation of the last vowel.[2] However, most of the rules

[2] See Plénat (1987) for a particularly elaborate and insightful attempt.

that are identified in these cases are unconvincing, for various reasons.

First, the variation observed has different degrees of generality in the language, and may correspond to more or less natural phonological processes. If we go back to the examples of Table 1, while diphthongation can is a natural and widespread phonological phenomenon, the alternation between [ə] and [jɛ] is uncommon in French, and only observed in a handful of verbs. On the other hand, an alternation between a vowel and a V+[l] sequence is quite common in French (the [u]/[ɔl] alternation is found, for instance, in some adjectives, like *fou*$_{\text{MSG}}$ / *folle*$_{\text{FSG}}$ 'mad', *mou*$_{\text{MSG}}$ / *molle*$_{\text{FSG}}$ 'soft'), but can hardly described as a natural phonological process.

In addition to such cases of segmental alterations, variation in inflectional paradigms often involves the adjunction of affix-like sequences (traditionally called 'augments' or 'stem extensions') whose phonological structure is, by definition, unpredictable; the two types of phenonena more often than not occur simultaneouslu. Tables 2a and 2b provide some examples of verbal paradigms in (Central) Catalan and in Italian[3].

Person					
1SG	2SG	3SG	1PL	2PL	3PL
'donu	'donəs	'donə	du'nɛm	du'nɛw	'donən
səɾ'βɛʃu	səɾ'βɛʃəs	səɾ'βɛʃ	səɾ'βim	səɾ'βiw	səɾ'βɛʃən

TABLE 2a. PRESENT INDICATIVE OF THE CATALAN VERBS DONAR ('GIVE') AND SERVIR ('SERVE')

Person					
1SG	2SG	3SG	1PL	2PL	3PL
'tɛngo	'tjɛni	'tjɛne	te'njamo	te'nete	'tɛngono
fi'nisko	fi'niʃʃi	fi'niʃʃe	fi'njamo	fi'nite	fi'niskono
ko'nosko	ko'noʃʃi	ko'noʃʃe	konoʃ'ʃamo	konoʃ'ʃete	ko'noskono
'ɛsko	'ɛʃʃi	'ɛʃʃe	uʃ'ʃamo	uʃ'ʃite	'ɛskono

TABLE 2b. PRESENT INDICATIVE OF ITALIAN VERBS TENERE ('KEEP'), FINIRE ('FINISH'), CONOSCERE ('KNOW') AND USCIRE ('GO OUT')

[3] In Tables 2a-2b we adopt a broad phonological transcription which leaves aside some specific processes, such as vowel lengthening in stressed position in Italian. More specifically, in Catalan the [o]/[u] alternation observed in the paradigm of DONAR can also be interpreted as the outcome of a regular phonological rule of vowel reduction in unstressed position, although the identification of an underlying /o/ is not straightforward in all cases (cf. Guerrero 2011); in Italian, [ʃ] is intrinsically geminated, [j] is regularly deleted after a palatal (like in [konoʃ'ʃamo], [uʃ'ʃamo]), the [ɛ]/[e] alternation is not taken into account, as the transcriptions proposed are based on a Central Italian standard pronunciation, although the concrete realisations of these phonemes are extremely variable in different varieties of spoken Italian.

Both Catalan and Italian display, for some verbs, the adjunction of a segment preceding the person-marking ending in some cells of the paradigm. However, while this segment is constant in Catalan ([ɛʃ]), in Italian it apparently undergoes a process of palatalisation ([isk]/[iʃʃ]) which is also observed in other verbs like CONOSCERE or USCIRE. Italian verbs, in particular, show how (more or less natural) phonological processes and non-phonological processes may combine in various ways to determine the final shape of inflected word forms, thus casting serious doubt on the possibility of proposing a unitary phonological explanation for the observed variation. [4] Table 3 shows the different types of variation found in the inflection of the four Italian verbs listed in Table 2b, and their combinations[5].

Person					
1SG	2SG	3SG	1PL	2PL	3PL
g-I	D	D	–	–	g-I
EXT	EXT+P	EXT+P	–	–	EXT
–	P	P	P	P	–
–	P	P	VM+P	VM+P	–

TABLE 3. TYPES OF VARIATION FOUND IN THE PARADIGM OF SOME ITALIAN VERBS

Tables 1, 2a-b also show that all the variation is distributed in the same way in each language, independent of its type and of its phonological naturalness. Table 4 shows the distribution of allomorphic variants in the present indicative of French, Catalan and Italian, as it emerges from the data presented above[6].

	Person					
	1SG	2SG	3SG	1PL	2PL	3PL
French	A	A	A	B	B	C
Catalan	A	A	A	B	B	A

[4] See Aronoff (2012) for a similar argumentation based on French.

[5] The abbreviations used are: g-I = [g] (velar) insertion; D = diphthongation; EXT = extension (insertion of an extra segment); P = palatalisation; VM = vowel modification. In order to list the modifications in Table 3, we selected as 'basic stem' the one that allows obtaining all the forms with the least number of modifications (respectively, 1-2PL for TENERE and FINIRE, and 1SG-3PL for CONOSCERE and USCIRE), but it is clear that it is an arbitrary choice. Choosing another form, however, would have made the picture more complex, but would not have changed the gist of our discussion. In particular, for recent syntagmatic accounts of Italian verbal inflection see e.g. Burzio (2003, 2004); see also Thornton (2007) for some observations on paradigmatic vs. syntagmatic approaches.

[6] Catalan displays, in fact, a slightly more complex pattern, since 1SG may be different from the rest of singular (see Guerrero 2011 for a complete overview).

Italian	A	B	B	C	C	A

TABLE 4. DISTRIBUTION OF VARIATION IN THE PRESENT INDICATIVE OF FRENCH, CATALAN AND ITALIAN VERBS

We will go back to stem distribution later. What counts now is to observe that there is a compelling explanation for the observed variation and its distribution: it directly reflects the various historical changes the relevant languages have gone through. Consequently, any attempt to explain this variation as part of the synchronic grammar amounts to a synchronic recapitulation of historical changes. Since the historical changes already motivate why the systems are shaped the way they are, one may doubt that there is anything to gain from this synchronic projection: in the absence of empirical studies showing that speakers are aware of the fine-grained morphophonological structure of table 3, rather than just the purely morphological organisation of table 4, the simpler theory is the one that postulates a simpler synchronic analysis (see Maiden 2005, Pirrelli & Battista 2000 for some relevant observations).

3. A THEMATIC APPROACH TO ROMANCE VERBAL PARADIGMS

3.1 The formal representation of lexemes

Several studies in theoretical morphology and in psycholinguistics converge today in considering that morphological competence cannot be reduced to a binary distinction between what is irregular and memorised and what is regular and produced on-line. Rather, it is quite commonly admitted that frequent regular forms are memorised by speakers just as commonly as irregular forms (cf. Stemberger & MacWhinney 1988; Baayen *et al.* 1997, 2003). Moreover, the assumption that the added compactness of the lexicon ensured by a constructive approach is virtuous only holds if one assumes (i) that there are strong memory limitations, and (ii) that lexical memory is more costly than lexical processing. Nothing in the relevant literature suggests that this is true.

Once we acknowledge that lexemes are complex objects which may be associated with multiple phonological representations, the task of an abstractive model, is to achieve economy of description not by reducing the amount of memorised information, but by finding means of organising this complexity. Such theoretical objects as morphomes (Aronoff 1994; Maiden 2005, 2009) and stem spaces (Bonami & Boyé 2003, 2007; Montermini & Boyé 2012) have precisely the goal of structuring paradigms according to

purely morphological principles, independent of external (e.g. phonological or semantic) motivations.

3.2 Units of morphological analysis

The most neutral way of addressing the question of paradigmatic relations is the one expressed by Ackerman *et al.* (2009) (see also Malouf & Ackerman 2010) as the "Paradigm Cell Filling Problem" in the following terms:

> Given exposure to an inflected wordform of a novel lexeme, what licenses reliable inferences about the other wordforms in its inflectional family?

In principle, a lexeme's paradigm may be entered from any of its inflected forms and thus the content of any cell should be inferrable from the content of any other cell. For complex paradigms like those of Romance verbs, which contain about fifty cells, this gives some 2,450 connections that are potentially relevant in order to fill all their cells. However, most of these connections are trivial, because some families of cells covary systematically, forming what Bonami & Boyé (2003) call *zones* and Ackerman and al. (2009) call *alliances of forms*. As we will see, theoretical description and experimental analysis converge in identifying these subdivisions of the paradigmatic space and their organisation.

An advantage of abstractive approaches is that are not committed any *a priori* position about the number and the nature of the pertinent units that are supposed to emerge from the analysis besides words. For constructive approaches, the basis of morphological analysis is the construction of forms from smaller units and the identification of rules that describe all and only the correct combinations. In abstractive approaches, on the other hand, stress is put on the relations between forms, with no presupposition on the way in which these relations can be expressed. In fact, they can be expressed in terms of a phonological string adjoined at the margins of another form, but they can also correspond to another kind of relation, for instance a non-segmental operation, an internal modification, and even an identity relation (like, for instance, between persons 1SG, 2SG and 3SG of the French verbs illustrated in Table 1).

The verbal systems of Romance languages, in general, share two main characteristics: a strong redundancy in the endings, even across verbal classes; and a great deal of internal variation across paradigms. In this, Romance verbs distinguish themselves from what we may consider more prototypical examples of inflection class variation, like for instance Latin noun classes. In Latin noun declension, variation is exclusively located at the

right margin of word forms, and patterns differ in a significant way, with little reuse of the same exponents across classes. An analysis of paradigm cell relations of Romance verbs that does not establish *a priori* how many and which units we should find, leads naturally to the identification of word form exponents as the maximal rightmost strings that are common across the patterns, and to interpret all the remaining variation as stem allomorphy.[7] Consequently, all remaining elements are not considered synchronically contentful, but result from a particular relation between stems. This allows for a reduction of such elements as theme vowels or stem extensions, whose status as contentful units is particularly doubtful. The meaning of theme vowels does not correspond to a morphosyntactic value, but rather to a purely morphological feature, i.e. the inflection pattern to which a particular verb belongs. However, they play this role only for a part of the forms, since paradigms contain ambiguous forms, where theme vowels do not appear on the surface. In addition, the inflection pattern of a lexeme may be inferred from a sequence which is not necessarily placed between the stem and the ending (the locus of traditional theme vowels in Romance), as shown, for instance, by such French verbs like PESER ('weight' $[pɛz]_{1SG\ PRES\ IND}$ / $[pəzɔ̃]_{1PL\ PRES\ IND}$) vs. SCELLER ('seal' $[sɛl]_{1SG\ PRES\ IND}$ / $[selɔ̃]_{1PL\ PRES\ IND}$), or by Italian verbs with so-called mobile diphthongs (cf. van der Veer 2006; Montermini & Boyé 2012: 79). Such partial predictors of inflection class are thus doubtful morphological units both from the point of view of form and from the point of view of meaning; treating them as parts of a stem avoids having to raise any of these issues.

3.3 The Stem space

A second observation is that stem variation is distributed, across paradigms, in a systematic way. Although various attempts have been made to account for this variation on non-morphological grounds, for instance semantic or phonological, it is quite commonly admitted today that the majority of the stem allomorphies found in Romance conjugation cannot receive a synchronic explanation, and should be interpreted as the outcome of a purely morphological, or morphomic (Aronoff 1994) distribution. Stems are generally formally defined as the phonological material on which a word form is built. Here, we propose a definition of stem which is both formal and distributional: in an inflectional paradigm, a stem is the minimal common

[7] This strategy is due to Boyé (2000). See Spencer (2012) and Bonami & Boyé (in press) for relevant discussion.

substring of a set of forms that are in systematic co-variation.[8] The irreduceable phonological information associated with a lexeme consists of a collection of stemsof. Stems are indexed, and languages specify with which stem a specific cell in the paradigm should be filled. In other terms, for a given class of lexemes a language specifies a stem space, corresponding to the set of procedures that allow to select the appropriate stem for each cell in the paradigm. Visually, we represent stem spaces as paradigmatic grids, in which each cell contains a stem index. Table 5 shows the Stem space of Italian verbs.[9]

<table>
<tr><td></td><td colspan="6">Person</td></tr>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr>
<tr><td>Future indicative</td><td colspan="6" rowspan="2">S6</td></tr>
<tr><td>Present Conditional</td></tr>
<tr><td>Present Subjunctive</td><td colspan="3">S2</td><td colspan="2">S4</td><td>S2</td></tr>
<tr><td>Present Indicative</td><td></td><td colspan="2">S3</td><td></td><td></td><td></td></tr>
<tr><td>Imperfect Indicative</td><td colspan="6" rowspan="2">S1</td></tr>
<tr><td>Imperfect Subjunctive</td></tr>
<tr><td>Preterite Indicative</td><td>S5</td><td></td><td>S5</td><td></td><td></td><td>S5</td></tr>
<tr><td>Imperative</td><td></td><td>S3</td><td></td><td>S4</td><td></td><td></td></tr>
<tr><td>Present Participle</td><td colspan="6">S1</td></tr>
</table>

[8] This definition is slightly idealised in not allowing for nonconcatenative inflectional operations; see Bonami & Boyé (in press) for a more careful statement.

[9] This stem space is an evolution of the one already presented in Montermini & Boyé 2012 which, in turn, derives from Pirrelli & Battista's (2000) 'Overall Distribution Schema'. As in Pirrelli and Battista's analysis, the Stem space we propose holds for all Italian verbs apart from eight highly irregular verbs (ANDARE 'go', AVERE 'have', DARE 'give', DIRE 'say', ESSERE 'be', FARE 'do', SAPERE 'know', STARE 'stay'). The integration of these verbs would have given a slightly more complex grid, but would not have affected the substance of our analysis. Concerning imperative, we consider that only 2SG, 1PL and 2PL are 'genuine' imperative forms in Italian, since they are the only ones which can take an enclitic pronoun (*mangialo*, *mangiamolo*, *mangiatelo*, cf. Graffi 1996). By comparison with Table 4, present indicative contains an extra cell for 1PL (here labelled S4). The identification of this stem as autonomous is only justified for two verbs in Italian, DOVERE (cf. *dobbiamo*$_{1PL\ PRES\ IND}$ vs. *dovete*$_{2PL\ PRES\ IND}$), whose paradigm presents several inflectional particularities (including overabundance of forms in [bb] and in [v], cf. Thornton 2012: 185) and DOLERE (cf. *dogliamo*$_{1PL\ PRES\ IND}$ [doʎ'ʎa:mo] vs. *dolete*$_{2PL\ PRES\ IND}$ [do'le:te]), although the distinction between the phoneme [ʎ] and the sequence [lj] is weak.

Gerund	
Past Participle	S7
Infinitive	S8

TABLE 5. THE STEM SPACE OF ITALIAN VERBS

An important property of the model presented here is that it is agnostic as to lexical storage. The model exhibits the internal organization of a paradigm in a way that makes it possible to infer unknown forms, but is perfectly compatible with the memorization of regular forms. Thus the identification of a stem space for the lexemes of a given language allows to achieve a first level of simplification: eight stems are sufficient to fill a paradigm containing fourty-nine cells (if we only consider simple – i.e. non-periphrastic – forms); in other words, a fluent speaker knowing one form based on each of these eight stems for a given lexeme has sufficient knowledge to solve the Paradigm Cell Filling Problem with certainty – although in practice he may very well know more than 8 forms.
Moreover, the distribution of these stems within the paradigm is highly constrained: in fact only 29 of all the possible combinations of eight stems in a fourty-nine cells paradigm are attested by actual Italian verbs.

3.4 The stem graph

Table 5 illustrates the maximal complexity that can theoretically be achieved by an Italian verb (with the exception of the eight listed in note 4). In practice, there is no attested verb displaying such a degree of complexity, and the most complex of Italian verbs, DOLERE (but cf. note 4), requires six stems to be memorised. In fact, for the majority of verbs, knowing one stem is sufficient, under the assumption of default strategies used to inflect regular verbs. Under this view, the relations connecting a stem to the forms it constructs and the relations connecting stems between them are of the same nature and may be expressed by the same formalism. Irregularity may then be viewed as a deviation from the default expectation that needs to be explicitly stated. This can concern either the relation between a stem and a form, or the relation between two stems, leading respectively to inflected form suppletion, or to unpredictable stem alternation.

In order to exemplify how the model works, let us go back to the Italian examples. On the basis of the most numerous (sub)classes of Italian verbs, we identify some default stem-to-form relations. The ones holding for the present indicative are listed in Table 6.

Person

1SG	2SG	3SG	1PL	2PL	3PL
S2	S3	S3	S4	S1	S2
XV–Xo	XV–Xi	X–X	X–Xjamo	X–Xte	Xi–Xano Xa–Xono

TABLE 6. DEFAULT STEM-TO-FORM RELATIONS FOR THE PRESENT INDICATIVE OF ITALIAN VERBS

The formalism adopted in Table 6 should be read as follows:
- 1SG is constructed on S2 by replacing its final vowel with [o];
- 2SG is constructed on S3 by replacing its final vowel with [i];
- 3SG is identical to S3;
- 1PL is constructed by adding the sequence [jamo] to S4;
- 2PL is constructed by adding the sequence [te] to S1;
- 3PL is constructed by adding the sequence [no] to S2 and by replacing its final vowel by [a] if it is [i] and by [o] if it is [a].

The first two relations, in particular, capture the fact that all 1SG PRES IND and 2SG PRES IND share the same endings in Italian, independent of the inflectional class, while the last set of relations accounts for the fact that for a class of verbs (those whose infinitive ends in *-are*, like LAVARE 'wash') the 3SG PRES IND ends in [ano] (*lavano*) and that for other (sub)classes (those whose infinitive ends in *-ere* or *-ire* like TEMERE 'fear' or DORMIRE 'sleep') it ends in [ono].

In addition to the relations illustrated in Table 6, the system specifies a series of default relations between stems. It should be noted, however, that what exactly holds as a default relation is an empirical question. Let us take a couple of examples. For whose infinitive ends in *-are*, which constitute the most numerous and productive class (all neologisms and adapted borrowings, for instance, belong to this class), stem are distinguished only by vowel variations at the right margin, as illustrated in Table 7 below (only the relations relevant for present indicative are listed). Of the 2,073 verbs contained in the LIP (*Lessico di frequenza dell'italiano parlato*, De Mauro *et al.* 1993) corpus 1,504 (72,5%) belong to this class.

S1–S2	S1–S3	S1–S4	S2–S3	S2–S4	S3–S4
Xa–Xi	X–X	XV–X	Xi–Xa	XV–X	XV–X
lava–lavi	lava–lava	lava–lav	lavi–lava	lavi–lav	lava–lav

TABLE 7. STEM-TO-STEM RELATIONS FOR THE PRESENT INDICATIVE OF ITALIAN CLASS-1 (INFINITIVE IN *-ARE*) VERBS

Right-vowel variation also characterises stem relations for other classes of verbs. In particular, there are at two other subclasses for which stem-to-stem relations exclusively correspond to vowel variations (Table 8).

S1–S2	S1–S3	S1–S4	S2–S3	S2–S4	S3–S4
Xe–Xa Xi–Xa	XV–Xe	XV–X	Xa–Xe	XV–X	XV–X
teme–tema	teme–teme	teme–tem	teme–tema	tema–tem	teme–tem
dormi–dorma	dormi–dorme	dormi–dorm	dorma–dorme	dorma–dorm	dorme–dorm

TABLE 8. STEM-TO-STEM RELATIONS FOR THE FOR THE PRESENT INDICATIVE OF TWO ITALIAN SUBCLASSES (INFINITIVE IN *-ERE* AND *-IRE*) VERBS

Interestingly, lexemes exhibiting this pattern are a minority in their respective classes (traditional Class 2 and 3): in the LIP, only 26 of 381 verbs ending in *-ere*[10] and 32 of 206 verbs ending in *-ire* behave like TEMERE and DORMIRE respectively. The majority of verbs traditionally included in these two classes display more than a mere vowel change in their stem relations. On one side, the large majority of *-ire* verbs (135 of 206) display a stem extension [isk]/[iʃ] and behave like FINIRE illustrated in Table 2b. The situation for the *-ere* verbs is more sparse: the 381 verbs belonging to this set all exhibit the alternations illustrated in Table 8, but for several of them these are combined with further modifications. Overall these verbs may be divided in 64 different subclasses; some phonological phenomena, such as the palatalization of velar stops (cf. *torco*$_{1SG\ PRES\ IND}$ ['tɔrko] / *torci*$_{2SG\ PRES\ IND}$ ["tɔrtʃi] 'twist', *spargo*$_{1SG\ PRES\ IND}$ ['spargo] / *spargi*$_{2SG\ PRES\ IND}$ ['spardʒi] 'spread'), are recurrent, but may be combined with other phenomena in different ways. Thus, despite the fact that few verbs pattern exactly like those in Table 8, the relevant vowel alternations are the only common caracteristic holding the classes together.

Overall, the 2,073 verbs of the LIP define 74 different patterns of stem combinations. In Table 9 we list the eight sets of verbs which are represented by more than 15 lexemes, only indicating, for the sake of the present discussion, S1, S2, S3 and S7 (for which see below). (This table is both an expansion and a simplification of the one already presented in Montermini & Boyé 2012: XX; figures include derived verbs).

[10] This figure includes 39 verbs whose infinitive actually ends in *-rre* (e.g. PORRE 'pose') which, in all other respects, behave like *-ere* verbs. Moreover, if we extend stem variation to stress, the only *-ere* verb for which all the relations in Table 8 hold without exceptions is TEMERE, since its infinitive is paroxytonic like that of *-are* and *-ire* verbs (cf. [te'me:re] and [la'va:re]), whereas all other verbs of this subclass present at least the particularity of having a proparoxytonic stress (cf. *credere* 'believe$_{INF}$' ['kredere]).

S1	S2	S3	S7	total (LIP)	example
Xa	Xi	Xa	Xat(o)	1,504	LAVARE 'wash'
Xi	Xiʃʃe	Xiska	Xit(o)	135	FINIRE 'end'
Xi	Xe	Xa	Xit(o)	32	SENTIRE 'feel'
Xnde	Xnde	Xnda	Xz(o)	32	PRENDERE 'take'
Xdʒe	Xdʒe	Xga	Xt(o)	30	SPINGERE 'push'
Xde	Xde	Xda	Xz(o)	29	CHIUDERE 'close'
Xne	Xne	Xnga	Xst(o)	18	PORRE 'pose'
Xette	Xette	Xetta	Xess(o)	17	METTERE 'put'

TABLE 9. THE EIGHT MORE FREQUENT PATTERNS IN ITALIAN CONJUGATION

Deciding which patterns function as the default and which ones correspond to 'irregular' verbs should be seen as an empirical question, which can be answered through behavioral and/or psycholinguistic studies (cf. Bonami *et al.* for French, Giraudo *et al.* 2012 for Italian). What is important about the current model is that it allows an explicit definition of what a fully regular and a fully irregular lexeme correspond to. A regular lexeme is one for which all stems are linked by predictable relations, an irregular one is one for which at least one stem is not linked to the others by such a relation. So, for a speaker of Italian it is sufficient to memorise one stem in order to inflect a verb like LAVARE, while for such verbs like CONOSCERE 'know' or NASCERE 'be born' (which have a unique inflection pattern) they must memorise more than one stem to competently inflect the lexeme. Note again that the memorisation of one stem is *sufficient* for the inflection of a lexeme, which does not entail that other stems can not be memorized. What is actually memorised by each speaker and under which form is an orthogonal question to the formal characterisation of verbal inflection.

To sum up, the solution we propose for the Paradigm Cell Filling Problem is that individual cells are linked to one another indirectly, via stems that correspond to sets of cells in systematic co-variation. Moreover, stems do not distribute randomly in paradigms, but according to recurrent patterns, which allows an important reduction of the range of possible paradigms. Let us go back to the Italian verbal system. In theory, 28 connections are necessary in order to fill an eight-stem space. In fact, most of these connections too are redundant. In (1) we give the most general connections linking three stems in the Italian paradigm (only the relevant relations are indicated), whereas Table 10 gives the corresponding stems for four Italian verbs[11], where solidary stems are marked in grey.

[11] S7 is the stem on which past participle is constructed. Of course, past participle has itself

(1) S2–S3 Xi–Xa
Xa–Xe
S2–S7 Xi–Xat(o)
Xa–Xut(o)
S3–S7 Xa–Xt(o)
Xe–Xut(o)

	S2	S3	S7
LAVARE (‘wash’)	lavi	lava	lavato
CONOSCERE (‘know’)	konoska	konoʃʃe	konoʃʃuto
MORDERE (‘bite’)	mɔrda	mɔrde	mɔrso
NASCERE (‘be born’)	naska	naʃʃe	nato
*	A	B	A

TABLE 10. S2, S3 AND S7 IN FOUR ITALIAN VERBS

The four configurations illustrated in Table 10 are the only attested for all Italian verbs. What these data show is that S3 may be linked to S2, to S7 or to both, but that S2 ans S7 are linked only if they are also linked to S3, so that there is no existing Italian verb illustrating the configuration in the last row. We consider then that the S2-S3 and S3-S7 connections have a higher predictive value than the S2-S7 connection, and that these two stems should be linked only indirectly, via S3. This amounts to saying that if a speaker knows S2 and S3 for some lexeme and tries to infer S7 has no predictive advantage over a speaker who knows just S3, whereas they have an advantage over a speaker who knows just S2.

When we transfer this kind of analysis to the scale of entire paradigms, only nine of the 28 possible connections for the eight Italian stems appear to be predictive, thus allowing us to draw a graph in which only the stems which are good predictors of each other are connected. Figure 1 gives a graph showing the distribution of stems of Italian verbs and one showing the relations holding for a first conjugation verb (see Montermini & Boyé 2012 for a more detailed discussion, and Pirrelli & Battista 2000 for a similar proposal of stem-dependency trees), while Figure 2 shows the graphs of the two irregular verbs CONOSCERE and NASCERE.

four forms corresponding to the crossing of the masculine/feminine and singular/plural values; we put the ending corresponding to the masculine singular form in brackets.

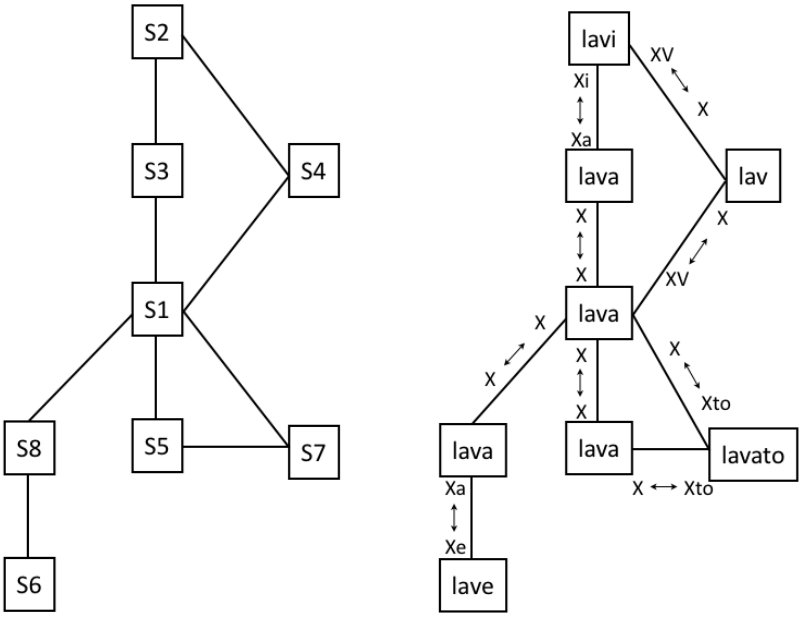

FIGURE 1: STEM GRAPH FOR THE ITALIAN FIRST-GROUP VERBS

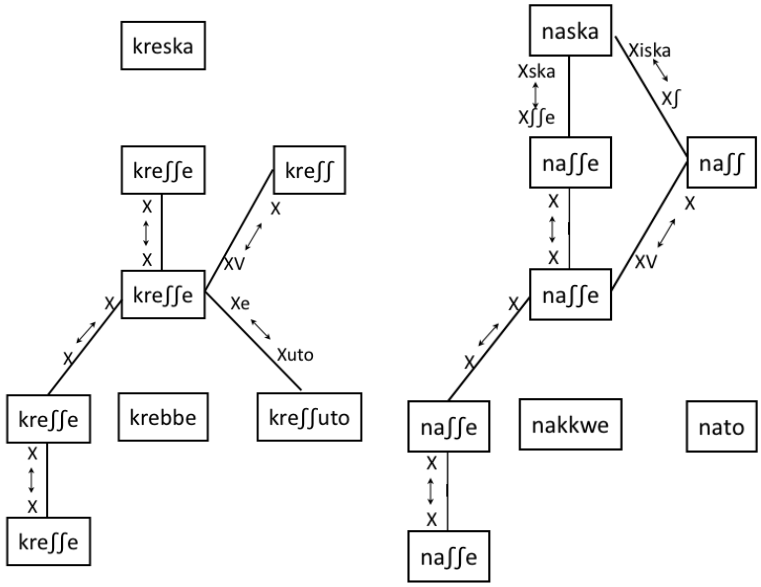

FIGURE 2: STEM GRAPHS FOR TWO ITALIAN IRREGULAR VERBS

Reducing paradigms to stem graphs like the ones we propose has several advantages. First, it allows for the identification of the most reliable implicative relations (Wurzel, 1984) between the morphomic stems that distribute throughout a paradigm. Second, it indicates exactly the number of forms (stems) that must minimally be associated with a lexeme in a speaker's memory under an assumption of default regular inflection: in the case of regular lexemes, like the verb LAVARE, one form is sufficient, since all the stems are connected by predictable relations; in the case of irregular lexemes, each portion of the graph which is not connected to the rest represents an extra stem that needs to be stored in the speaker's lexicon. Moreover, each connection can be represented by one or more relations, that can be mutually exclusive, complementary (like the ones illustrated in Tables 7-8 above and in (1)) or hierarchically ordered, capturing the fact that the classification of lexemes into classes and subclasses is more a gradient than a categorical matter. Finally, as we saw above, it allows reducing the number of connections keep track of to describe an inflection system: in a first step, we reduced the 2,352 possible relations between the 49 cells of the paradigm of an Italian verb to 28 symetric relations between stems and 49 relations from a cell to the stem it is based on. In a second step, we reduced the 28 symetric relations between the eight stems of an Italian verb to 9 on the basis of the relative interpredictibility between stems.

4. CONCLUSION

In this paper we argued in favor of an abstractive model of morphology, in which morphosyntactic properties are not expressed by subparts of complex word forms, but rather by fully inflected forms. Under this view, the main task of morphological analysis is not to identify the rules for combining minimal elements, but to describe the relations between forms and the generalisations that can be drawn from these relations. An abstractive model is compatible with an enriched view of the lexicon, in which lexemes may have complex entries containing multiple phonological representations. This is coherent with psycholinguistic evidence that speakers memorise at least frequent regular forms along with unpredictable irregular forms. In this framework, economy of description is not achieved by reducing the amount of memorised information, but by giving an explicit description of how this information is organised and is the basis of inferences made by speakers. Stem spaces and stem graphs are means for constraining variation and organising complexity. As we observed above, the question of what exactly

is encoded in speakers' brains is an issue orthogonal to formal treatments, such as the one we proposed here. Our proposals aim at defining the limits of morphological competence. The knowledge actually encoded in each speaker's brain may vary, but only varies within these limits.

The approach outlined here bears a family resemblance with other, more recently developed approaches to the implicative structure of paradigms, most notably Finkel & Stump's (2007, 2009) work on principal parts and the information-theoretic approach of Ackerman *et al.* (2009). Let us quickly outline how these approaches differ. Finkel & Stump's work focusses on the identification of collections of paradigm cells that allow for categorical inference of the rest of the paradigm, for all lexemes not taking into account any notion of (ir)regularity or exceptionality. By contrast, the current approach relies centrally on the identification of the regular relations between proper subparts of the paradigm, as links in the stem graph express default implicative relations that irregulars may violate. We submit that this is a preferable design feature for a performance-compatible model. First, it entails that regular lexemes necessitate less lexical storage than irregulars – whereas under Finkel & Stump's approach, deeply irregular lexemes typically require fewer principal parts, precisely because a single exceptional form allows one to predict the rest of the paradigm. Second, it is congruent with undisputed observations about speaker behavior: in situations of uncertainty speakers assume lexemes to be regular, showing that there is some awareness of degrees of regularity.

Ackerman *et al.*'s approach is in principle much closer to the approach defended here. In their view implicative relations between two cells are evaluated by computing the conditional entropy of one cell given knowledge of another cell. The limiting case where a collection of cells are all related by implicative relations with null entropy are direct analogues of our notion of an indexed stem. Default relations in the stem graph then correspond to pairs of cells related by nonzero but comparatively low conditional entropy. Despite these similarities, there are important differences. Ackerman *et al.*'s approach is more general than the one defended here in at least three respects.

First, unlike the current approach, it does not postulate a division between implicative relations at the level of words and at the level of stems. Although this is mostly innocuous when dealing with Romance conjugation, it is more appropriate to the study of inflectional systems where the implicative structure lies in the distribution of exponents (see Sims 2010), or on properties of whole words that are not reducible to a segmented part (Blevins 2006).

Second, Bonami & Boyé (in press) show in the case of French

conjugation that the purely word-based assumptions of Ackerman *et al.* give rise to a slightly different picture of implicative relations than the stem-space analysis of Bonami & Boyé (2003). This is due to the fact that, while it is quite permissive as to the number of distinct stems it allows for a lexeme, the current approach still presupposes a minimal amount of segmentation of forms into stems and exponents. This leads to treating differently two kinds of deep irregularity: lexemes with suppletive inflected forms, which do not exhibit any regular exponent, and lexemes with exceptional stem selection patterns, which use an unexpected stem in some cell of their paradigm. Bonami & Boyé (in press) argue that this amounts to a trace of constructive *ethos* within an otherwise abstractive approach.

Third, Ackerman *et al.*'s approach relies on a quantitative notion of generality (entropy is computed on the basis of knowledge of the probability distribution of different implicative patterns) rather than on a categorical distinction between regular and irregular. Although it is an empirical question whether regularity is categorical or gradient, and that question can only be sorted out by minute psycholinguistic examination, the entropy-based approach allows one to use frequency distributions as a proxy for a regularity measure, and thus to make some progress in the absence of sufficient psycholinguistic data.

From this comparison it should be clear that the stem-based, default-theoretic approach presented here and the strictly word-based and information theoretic approach share a core of common assumptions on the design properties of inflection systems, differing mostly in the underlying modelling tools, and making very similar predictions. We submit that at this point in the history of the study of morphology, the stem and defaults approach still has advantages in terms of conceptual familiarity and readability of analyses, at least in the case of Romance conjugation, for which it was designed.

REFERENCES

Ackerman F., Blevins J.P. & Malouf R. (2009). Parts and wholes: Implicative structure in inflectional paradigms. In J. Blevins & J.P. Blevins (Eds.), *Analogy in Grammar* (pp. 54-82). Cambridge: Cambridge University Press.

Anderson, S. (1992). *A-morphous Morphology*. Cambridge: Cambridge University Press.

Aronoff, M. (1994). *Morphology by Itself*, Cambridge, MA: MIT Press.

Aronoff, M. (2012). Morphological stems: what William of Ockham really said. *Word Structure* 5: 28-51.

Baayen R.H., Dijkstra T. & Schreuder R. (1997). Singulars and plurals in Dutch:

Evidence for a parallel dual-route model. *Journal of Memory and Language* 37, 94-117.

Baayen R.H., McQueen J.M., Dijkstra T. & Schreuder R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In R.H. Baayen & R. Schreuder (Eds.), *Morphological Structure in Language Processing* (pp. 355-390). Berlin: Mouton de Gruyter.

Blevins J.P. (2006). Word-based morphology. *Journal of Linguistics* 42, 531-573.

Bonami O. & Boyé G. (2003). Supplétion et classes flexionnelles dans la conjugaison du français. *Langage* 152, 102-126.

Bonami O. & Boyé G. (2007). Remarques sur les bases de la conjugaison. In E. Delais-Roussarie & L. Labrune (Eds.), *Des sons et des sens. Données et modèles en phonologie et en morphologie* (pp. 77-90). Paris: Hermès-Lavoisier.

Bonami O. & Boyé G. (in press). De formes en thèmes. In F. Villoing, S. Leroy & S. David (Eds.), *Du régulier et du minutieux, Etudes de morphologie en hommage à Françoise Kerleroux*. Saint-Denis: Presses Universitaires de Vincennes.

Bonami O., Boyé G., Giraudo H. & Voga M. (2008). Quels verbes sont réguliers en français ? *Actes du premier congrès mondial de linguistique française*, 1511-1523.

Boyé G. (2000). *Problèmes de morphophonologie verbale en français, espagnol et italien*. PhD dissertation, U. Paris 7.

Brown D. & Hippisley A. (2012). *Network Morphology*. Cambridge: Cambridge University Press.

Burzio L. (2003). Output-to-output faithfulness: the Italian connection. *Lingue e linguaggio* II.1, 69-104.

Burzio L. (2004). Paradigmatic and syntagmatic relations in Italian verbal inflection. In J. Auger, J. Clancy Clements & B. Vance (Eds.), *Contemporary Approaches to Romance Linguistics* (pp. 17-44). Amsterdam: John Benjamins.

De Mauro T., Mancini M., Vedovelli M. & Voghera M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano: Etaslibri.

Derwing B.L. & Skousen R. (1988). Morphology in the mental lexicon: A new look at analogy. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 1989* (pp. 56-71). Dordrecht: Kluwer.

Graffi G. (1996). Alcune riflessioni sugli imperativi italiani. In P. Benincà, T. De Mauro, N. Vincent, G. Cinque (Eds.), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy* (pp. 133-148). Roma: Bulzoni.

Guerrero A. (2011). Verbal inflection in Central Catalan: a realisational analysis. *Lingue e linguaggio* X (2), 265-282.

Finkel, R. & Stump, G. (2007). Principal parts and morphological typology. *Morphology* 17, 39-75.

Finkel, R. & Stump, G. (2009). Principal parts and degrees of paradigmatic transparency. In J. Blevins & J.P. Blevins (Eds.), *Analogy in Grammar* (pp. 13-53). Cambridge: Cambridge University Press.

Hockett, C. (1954). Two models of grammatical description. *Word* 10, 210-234.

Jackendoff, R. (2002). *Foundations of Language. Brain, Meaning, Grammar and Evolution*. Oxford: Oxford University Press.

Maiden, M. (2005). Morphology autonomy and diachrony. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 2004* (pp. 173-175). Dordrecht: Kluwer.
Maiden M. (2009). From pure phonology to pure morphology. The reshaping of the Romance verb. *Revue de linguistique de Vincennes* 38, 45-82.
Malouf R. & Ackerman F. (2010). Paradigms: the low entropy conjecture. Paper presented at the Workshop on Morphology and Formal Grammar, Paris, July 8, 2010.
Montermini F. & Boyé G. (2012). Stem relations and inflection class assignment in Italian. *Word Structure* 5 (1), 69-87.
Pirrelli V. & Battista M. (2000). The paradigmatic dimension of stem allomorphy in Italian. *Rivista di linguistica* 12 (2) 38, 307-380.
Plénat, M. (1987). Morphologie du passé simple et du passé composé des verbes de l' "autre" conjugaison. *ITL Review of Linguistics* 77-78, 93-150.
Sims, A. (2010). Probabilistic paradigmatics: Principal parts, predictability and (other) possible pieces of the puzzle. Paper presented at the 14th International Morphology Meeting, Budapest.
Spencer, A. (2012). Identifying stems. *Word Structure* 5 (1), 88-108.
Stemberger J.P. & McWhinney B. (1988). Are inflected forms stored in the lexicon? In M. Hammond & M. Noonan (Eds.), *Theoretical Morphology. Approaches in Modern Linguistics* (pp. 101-116). San Diego: Academic Press.
Stump, G. (2001). *Inflectional Morphology*. Cambridge: Cambridge University Press.
Thornton A.M. (2007). Is there a partition in the present indicative of Italian regular verbs? *Annali online della Facoltà di Lettere e Filosofia dell'Università di Ferrara* II.2.
van der Veer B. (2006). *The Italian 'Mobile Diphthongs'. A Test Case for Experimental Phonetics and Phonological Theory*. Utrecht: Lot.
Wurzel, Wolfgang (1984). *Flexionsmorphologie und Natürlichkeit.* Berlin: Akademie-Verlag. Translated as Inflectional Morphology and Naturalness, Dordrecht: Kluwer, 1989.

Fabio Montermini
CLLE-ERSS, CNRS & Université de Toulouse
Université de Toulouse le Mirail
Maison de la Recherche
5, allées Antonio Machado
31058 Toulouse Cedex 9
France
e-mail: fabio.montermini@univ-tlse2.fr

Olivier Bonami
Université Paris-Sorbonne,
Institut Universitaire de France,
Laboratoire de Linguistique Formelle (CNRS & U. Paris Diderot)
1, rue Victor Cousin

75005 Paris
France
e-mail: olivier.bonami@paris-sorbonne.fr