Morphology (2023) 33:361–395
https://doi.org/10.1007/s11525-023-09414-7

Flexemes in theory and in practice

Modelling overabundance in Latin verb paradigms

Matteo Pellegrini[1]

Received: 3 January 2023 / Accepted: 28 June 2023 / Published online: 17 July 2023

Abstract
This paper provides an in-depth investigation of the possibility of systematically using flexemes – i.e., lexical units characterized in terms of form, as opposed to lexemes, characterized in terms of meaning – to model overabundance – i.e., the availability of more than one form in the same paradigm cell. The starting point is a preliminary evaluation of the advantages and disadvantages of using flexemes to account for different overabundance phenomena, showing that flexemes are a good way to capture the systematicity of overabundance, either across lexemes or across cells. Consequently, it is suggested that flexemes can be an interesting technical solution for the creation of a lexicon of Latin verbs that not only documents all the competing wordforms available as principal parts, but also captures the systematic relationship that sometimes holds between variants filling different cells. A principled method to identify such systematicity is then described in detail. It is argued that a constructive approach based on the identity of stems and/or inflection class is not fully adequate for the data at hand. Therefore, the proposed procedure adopts an abstractive, word-based perspective that only relies on alternation patterns between unsegmented wordforms. Practical and theoretical implications of the work are finally discussed, particularly regarding the usefulness of a formal approach to the identification of lexical units and paradigm cells.

Keywords Overabundance · Flexemes · Latin · Paradigm · Lexicon

1 Introduction: lexemes, flexemes and overabundance

The starting point of this paper is the discussion by Fradin and Kerleroux (2003; see also Fradin 2003: 235 ff.) of cases where a word formation process selects only one of the different meanings of a lexeme, be it a case of polysemy or homonymy. For

M. Pellegrini
matteo.pellegrini@unicatt.it

[1] CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milano, Italy

Table 1 Many-to-one mapping from lexemes to flexemes: Fr. CÉRÉBRAL ‘cerebral’

lexeme	flexeme	M.SG	F.SG	M.PL	F.PL
CÉRÉBRAL$_1$ CÉRÉBRAL$_2$	CÉRÉBRAL	*cérébral*	*cérébraux*	*cérébrale*	*cérébrales*

instance, the French adjective CÉRÉBRAL‘cerebral’ can be used either as a qualifying adjective – like in (1a) – or as a relational adjective – like in (1b). However, the derivative CÉRÉBRALITÉ ‘intellectuality’ only selects the qualifying meaning, as the contrast between (1c) and (1d) shows.[1]

(1) (a) *Il a opté pour une peinture cérébrale*
‘He chose a cerebral way of painting’
(b) *Le lobe cérébral a été atteint*
‘The cerebral lobe has been affected’
(c) *La cérébralité de sa peinture*
‘The intellectuality of his way of painting’
(d) **La cérébralité de ce lobe*
‘The cerebrality of this lobe’

Based on such cases, the authors argue that, if we want lexemes to be the input of word formation processes, as it is usually assumed, then they should be fully specified for a concrete meaning, rather than having only an abstract semantic representation encompassing all possible meanings. Hence, in this example, we would have (at least) two distinct lexemes, CÉRÉBRAL$_1$ with a qualifying meaning and CÉRÉBRAL$_2$ with a relational meaning, and the word formation process creating deadjectival nouns in *-ité* would apply to CÉRÉBRAL$_1$ but not to CÉRÉBRAL$_2$. Consequently, Fradin and Kerleroux (2003: 193) also introduce a different unit, that they call the “inflecteme”, or “flexème” in French (cf. also Fradin, 2003: 259),[2] that is underspecified for meaning and captures the identity of forms of different lexemes. In our example, the two lexemes CÉRÉBRAL$_1$ and CÉRÉBRAL$_2$ would map to the same flexeme CÉRÉBRAL, because all their wordforms are identical, as illustrated in Table 1.[3]

Shortly after the publication of this paper, interest in overabundance – defined as the availability of more than one inflected wordform (“cell mates”, cf. Thornton, 2019: 223) to express the same morphosyntactic property set for the same lexeme – began to grow in the literature on morphological theory. Thornton (2011) brought this phenomenon to the attention of theoretical morphologists, investigating it in the context of the canonical approach to typology (cf. Corbett, 2005, Brown et al., 2012,

[1] From Fradin and Kerleroux (2003: 178).

[2] In this paper, we will adopt “flexeme” as the English translation of Fr. “flexème”, following later works such as Thornton (2018) and Bonami and Crysmann (2018), as well as Fradin (2019: 71) himself.

[3] Fradin and Kerleroux (2003: 193) and Fradin (2003: 259) simply equate the notion of flexeme with the stem(s) of a lexical item; in a more recent development, Fradin (2019: 71) implies that we have distinct flexemes whenever the sets of wordforms of two lexical items are not identical, in a word-based perspective that we will adopt in this paper. See Thornton (2018: 304 f.) for more details on the history of the development of the concept of flexeme.

Table 2 One-to-many mapping from lexemes to flexemes: It. LENZUOLO 'bed sheet'

lexeme	flexeme	SG	PL	inflection class
LENZUOLO	LENZUOLO$_1$	*lenzuolo*	*lenzuoli*	1 (*-o*/*-i*)
	LENZUOLO$_2$	*lenzuolo*	*lenzuola*	5 (*-o*/*-a*)

Bond, 2019), where it can be considered as a deviation from the canonical expectation of uniqueness of realization of paradigm cells, and recently proposing a full-fledged canonical typology of overabundance (Thornton, 2019). Many other theoretical discussions (cf., among else, Stump, 2016; Guzmán Naranjo, 2019; Beniamine, 2021) and empirical studies (cf., among else, Bermel & Knittl, 2012, Cappellaro, 2013, Lečić, 2015; Pellegrini & Ricca, 2019; Guzmán Naranjo & Bonami, 2021) have appeared since then on this topic. Among them, the most relevant to this study are Thornton (2018) and Bonami and Crysmann (2018), that, in a recent reappraisal of the notion of flexeme, propose to apply it to the phenomenon of overabundance. For instance, the plural of the Italian noun LENZUOLO 'bed sheet' can be either *lenzuoli* – following inflection class 1 in the classification proposed by D'Achille and Thornton – or *lenzuola* – following inflection class 5 in that classification. Thornton (2013: 465 f.) shows that the two plural forms can be used interchangeably in (at least some of) the same contexts. Hence, in this case we have the mirror image of the situation shown in Table 1 above: rather than two semantically distinct lexemes that map to the same flexeme, this time there is a single lexeme LENZUOLO – as there is no change in meaning – that maps to two separate flexemes LENZUOLO$_1$ and LENZUOLO$_2$, each with its own inflection class and paradigm, although they share the same singular form, as summarized in Table 2.

In this paper, we explore the possibility of using flexemes to model overabundance in more detail. In Sect. 2, we preliminarily evaluate the advantages and disadvantages of using flexemes to model different types of overabundance, using Thornton's (2019) typology as our starting point, but also considering other distinctions that can potentially be relevant. We argue that the crucial point is that flexemes are useful to capture the systematicity of overabundance, either across lexemes or across cells. In Sect. 3, we add that flexemes can be an interesting technical solution for the creation of a lexical resource that lists the principal parts of Latin verbs, aiming not only at documenting the availability of different inflected wordforms in the same cell of a given lexeme, but also at capturing the systematic relationship that sometimes holds between variants filling different cells. In Sect. 4, we propose a principled method to identify such systematicity, showing that a constructive perspective (in the sense of Blevins, 2016: 14 ff.) based on the identity of stems and/or inflection class is not fully adequate for the data at hand, and thus proposing an abstractive (see again Blevins, 2016), word-based approach that relies on alternation patterns between unsegmented wordforms. Section 5 concludes the work by pointing out its practical and theoretical implications, particularly regarding the usefulness of a formal approach to the identification of lexical units and paradigm cells.

2 Flexemes and different types of overabundance

The idea of using flexemes to model overabundance is not new: it has been explicitly put forward in at least two studies, namely Thornton (2018) and Bonami and Crysmann (2018). However, both of these studies only argue that such a solution is adequate in specific cases of overabundance; whether this can be extended to all types of overabundance is left as an open question (cf. Bonami & Crysmann, 2018: 198;[4] Thornton, 2018: 312-313, 319, Fn. 14). Such a question is especially relevant as many different overabundance phenomena have been identified, and different types of overabundance have been shown to interact in a different way with the overall inflectional system, at least in some cases (cf. Guzmán Naranjo & Bonami, 2021). In this section, we address this question in more detail, using Thornton's (2019) canonical typology of overabundance as our starting point, and discussing whether overabundance phenomena that vary across the several dimensions identified by Thornton can be more or less satisfyingly treated by means of flexemes. In the framework of canonical typology, it is crucial to identify criteria against which actual examples of the phenomenon under investigation can be matched to see how close they are to a "canonical", indisputable case. Thornton (2019) identifies four such criteria for overabundance. Although we are less interested in how canonical different cases of overabundance are in this context, Thornton's criteria nevertheless provide a principled way to classify different specific examples, so we will exploit them to evaluate the usefulness of flexemes for different types of overabundance: they will be considered in detail in Sects. 2.1, 2.2, 2.3 and 2.4, before summing up and motivating our decisions in Sect. 2.5.

2.1 Presence of conditions

One of Thornton's criteria is related to the presence of conditions that influence the use of the available variants. Overabundance is considered to be more canonical if the cell mates are in free variation, and less canonical if there are conditions, be them sociolinguistic (diatopic, diastratic, diaphasic, or diamesic) or internal to the grammatical system (phonological, morphological, syntactic, semantic), with pragmatic conditions as a bridging case between these two kinds (cf. Thornton, 2019: 246). This criterion is arguably orthogonal to the issue of the usefulness of a modelling with flexemes. As flexemes are defined as purely formal objects that account for the inflectional behavior of a lexical item, they appear to be equally well-suited for cases of unconditioned and conditioned overabundance. The presence of conditions – especially semantic ones – only has repercussions on the choice to introduce different lexemes as well, alongside different flexemes.

For instance, let us consider the data of Table 3. This example is at first sight fully comparable to the one of LENZUOLO shown in Table 2 above. In both cases, in the plural both the form in *-i* and the form in *-a* are possible. The only difference is that

[4] On the other hand, Bonami and Crysmann (2018) sketch hierarchies of the generalizations that can be drawn on the systems of stem classes and inflection classes of French verbs and Czech nouns, respectively. This topic falls outside the scope of the present work – although, in principle, such hierarchies can be useful to account for the data considered here too.

Table 3 One-to-one mapping from lexemes to flexemes: It. MEMBRO 'member'/'limb'

lexeme	flexeme	SG	PL	Inflection class
MEMBRO$_1$ 'member'	MEMBRO$_1$	*membro*	*membri*	1 (*-o/-i*)
MEMBRO$_2$ 'limb'	MEMBRO$_2$	*membro*	*membra*	5 (*-o/-a*)

in the case of LENZUOLO, as we have seen above, the two forms can be used interchangeably, while in the case of MEMBRO, *membri* is used to refer to 'members' of an organization, while *membra* indicates the 'limbs' of the body: the corpus data provided by Thornton (2013: 462 ff.) show that the two plural forms are in complementary distribution between the two senses. Therefore, there is no doubt on the presence of distinct flexemes in both cases, on the basis of the presence of distinct wordforms in the plural and, consequently, different inflection class assignments. However, in the case of MEMBRO it seems reasonable to have also distinct lexemes, with a one-to-one mapping with the corresponding flexemes: we would thus have on the one hand a lexeme MEMBRO$_1$ meaning 'member' and mapping to the flexeme with plural *membri*, belonging to class 1, on the other hand a lexeme MEMBRO$_2$ meaning 'limb' and mapping to the flexeme with plural *membra*, belonging to class 5; both (f)lexemes share the singular form *membro*, that can be used with both meanings. Conversely, in the case of LENZUOLO the two flexemes map to the same lexeme, given their semantic interchangeability (see again Table 2 above).

If the semantic conditioning is not categorical, but only preferential, as it often happens, then the mapping between lexemes and flexemes is less straightforward. In this paper, we do not investigate this issue, as we only focus on flexemes as purely formal units – whose introduction is warranted by the presence of distinct wordforms in some paradigm cells – rather than on the mapping with corresponding lexemes on the basis of their content.

Similar considerations can be made for sociolinguistic conditions: the presence of distinct wordforms is sufficient to justify the introduction of different flexemes; the presence of reliable associations between distinct forms and specific geographic provenances, social groups and/or stylistic aspects might be considered as an argument in favor of the introduction of different lexemes corresponding to each flexeme. However, while introducing distinct lexemes due to differences in meaning is relatively unproblematic (as lexemes are defined on semantic grounds), doing so for sociolinguistic reasons is less straightforward, and alternative options are available. In the case of diatopic conditions, for instance, if the association between forms and geographic areas is exceptionless, the conclusion can be drawn that a given form variant is only attested in a language variety, and the other one in another language variety. In that case, each language variety can be analyzed as having its own form variant, i.e., its own flexeme and lexeme. Even when associations are less reliable, lexemes could be considered as not being the ideal place to code sociolinguistic information: in that case, flexemes are a good candidate as the locus of coding of these facts.

Table 4 Frequency ratio between cell mates and status of overabundance

Frequency ratio	Cell mate A	Cell mate B	Status of overabundance
$0: \geq 1 \leq \infty$	0	$1\text{-}\infty$	No overabundance, B is categorical
n:m	n	m(>n)	B is favored: conditioned overabundance?
n:n	n	n	A and B are balanced: canonical overabundance
m:n	m(>n)	n	A is favored: conditioned overabundance?
$\geq 1 \leq \infty: 0$	$1\text{-}\infty$	0	No overabundance, A is categorical

2.2 Frequency ratio between the cell mates

Another canonicity criterion has to do with the quantitative strength of overabundance, measured by the ratio between the number of attestations of the two cell mates in a corpus, ranging from one-to-one, corresponding to a case of perfectly balanced and canonical overabundance, up to the other logical endpoint, zero-to-one (or more), as summarized in Table 4.[5] The latter possibility corresponds to cases where only one form is attested, hence there is no overabundance and, consequently, no distinct flexemes can be introduced. Except for this trivial consideration, also this criterion is arguably orthogonal to the issue of the use of flexemes. The attestation of distinct wordforms appears to be sufficient by itself to justify the introduction of distinct units to account for them, from the perspective of language theory. However, the more the competition between cell mates is balanced, the more the overabundance phenomenon can be presumed to be perceived by speakers, and the more solid the entrenchment of both flexemes.

2.3 Uniqueness of cells

The other criteria proposed by Thornton (2019) are related to the systematicity of overabundance, across lexemes on the one hand, across cells on the other hand. In both cases, the idea is that the most canonical instances are the most isolated and idiosyncratic ones – those where only one lexeme or cell are involved. The more overabundance is systematic across lexemes or cells, the less it is canonical. As far as systematicity across cells is concerned, overabundance can involve a single cell (e.g., only the plural cell in the Italian examples given in Table 2 and Table 3 above), a set of cells – be it morphosyntactically defined (e.g., past participle cells of both genders and numbers in Italian verbs like PERDERE, with PST.PTCP.M.SG *perduto/perso*, PST.PTCP.F.SG *perduta/persa*, PST.PTCP.M.PL *perduti/persi*, PST.PTCP.F.PL *perdute/perse*) or morphomically defined (e.g., the set of overabundant cells in Italian verbs like SEDERE, with PRS.ACT.IND.1SG *siedo/seggo*, PRS.ACT.IND.3PL *siedono/seggono*, PRS.ACT.SBJV.1/2/3SG *sieda/segga*, PRS.ACT.SBJV.3PL *siedano/seggano*) – or all cells (e.g., both the singular and the plural cell of Italian nouns like ORECCHIO/-A, with SG *orecchio/orecchia*, PL *orecchi/orecchie*).

[5] From Thornton (2019: 241).

A potential interaction of this criterion with the appropriateness of flexemes as a modelling strategy has already been suggested by Thornton herself, specifically regarding the extreme case of overabundance in a single cell. As Thornton (2018: 312) points out, the presence of non-identical wordforms in one cell is technically a sufficient condition to introduce separate flexemes. However, given the idiosyncratic nature of similar overabundance phenomena, Thornton (2018: 313) recognizes the alternative option of having "a single flexeme with a single exceptional, overabundant cell".

2.4 Uniqueness of lexemes

As for systematicity across lexemes, overabundance can involve one isolated lexeme (e.g., idiosyncratic overabundance in the Russian noun meaning 'year', with GEN.PL *let*/*godov*), a set of lexemes (e.g., the alternation between strong and weak past (participles) in English verbs such as HANG with PST(.PTCP) *hanged*/*hung*, SNEAK, with PST(.PTCP) *sneaked*/*snuck*, and others), or even all lexemes of the relevant lexical category (e.g., the alternation between forms in *-ra-* and *-se-* in the imperfect subjunctive of all Spanish verbs, for instance HABER 'have' with IMPF.SBJV.1/3SG *hubiera*/*hubiese*, IMPF.SBJV.2SG *hubieras*/*hubieses*, and so on).

Also in this case, if an overabundance phenomenon is so idiosyncratic that it only involves a single lexeme, one might prefer to treat it as an exceptional behavior that does not require the introduction of distinct flexemes. Furthermore, this time, modelling strategies alternative to the introduction of distinct flexemes are intuitively more appealing in the most systematic case of overabundance phenomena that involve all lexemes. This is related to the fact that in such cases we would have to introduce distinct flexemes for all lexemes, which is not a very attractive solution in terms of economy of the description.[6] A possible alternative solution that is envisaged by Thornton (2018: 313) would be to build overabundance in the definition of lexemes, i.e., stating that lexemes can form a specific cell (or set of cells) in different ways.

On this topic, it is useful to take into account another dimension of variation that Thornton (2019) discusses, although she does not consider it as part of the canonical

[6] However, this is not necessarily true if the adoption of flexemes is implemented in different, more subtle ways. For instance, Bonami and Crysmann (2018) sketch a formalization of the lexeme-flexeme distinction couched in the framework of Information-based Morphology (IbM). They propose to distinguish between two different kinds of identifiers, "lexeme identifiers" (LID) – that are used to individuate lexemes – and "paradigm identifiers" (PID) – that are used to individuate flexemes, showing that in some cases it can be beneficial to associate multiple PIDs – i.e., flexemes – to the same LID – i.e., lexeme. They also organize PID objects into a type-hierarchy, arguing that lexically underspecified PIDs can be used to model some overabundance phenomena. In the present context, it is important to note that this allows them not to have to postulate a new entity in the lexicon for each flexeme: a flexeme is defined as the combination of a type in the hierarchy – which may correspond to an inflection class or to a stem alternation pattern – and a collection of stems; the PID is only the place where this information is recorded in the lexical entry. If two flexemes have, e.g., different stems, this can simply be coded by listing those different stems, without any need to introduce separate symbols to name them. In such an approach, there is no cost in using flexemes to code systematic overabundance, as it is sufficient to use an underspecified flexeme type in lexical entries: the grammar will then automatically generate multiple subtypes of that lexically listed information. The interested reader is referred to Bonami and Crysmann's (2018) paper for further details and references.

Table 5 Paradigm Linkage theory and different ways of being overabundant

content cell < L,σ >	form cell < Z{IC}, τ >	realized cell < *w*,τ >
< 'milk',SUP.ACC >	< *muls-*,SUP.ACC >	< *mulsum*,SUP.ACC >
	< *mulct-*,SUP.ACC >	< *mulctum*,SUP.ACC >
< 'clean',PRS.ACT.INF >	< *terg-*{2nd},PRS.ACT.INF >	< *tergēre*,PRS.ACT.INF >
	< *terg-*{3rd},PRS.ACT.INF >	< *tergere*,PRS.ACT.INF >
< 'love',PRS.PASS.IND.2SG >	< *am-*{1st},PRS.PASS.IND.2SG >	< *amāris,*PRS.PASS.IND.2SG >
		< *amāre,*PRS.PASS.IND.2SG >

typology. Thornton distinguishes different ways in which overabundance can arise, framing the issue in terms of Paradigm Linkage theory (Stump, 2016). This theory is based on the central claim that the definition of the inflectional morphology of a language consists in a mapping between three levels, namely from content cells – i.e., pairings of a specific lexeme (L) and a specific morphosyntactic property set (σ) – to realized cells – i.e., pairings of a specific wordform (w) and the morphosyntactic property set that it expresses (τ) – going through form cells – i.e., pairings of a specific stem (Z), possibly belonging to a specific inflection class (IC), and the relevant morphosyntactic property set (τ). The three levels are schematically expressed in the headers of Table 5.

Within this background, Thornton (2019: 229 ff.) notes that overabundance can arise either in the mapping from content cells to form cells, or in the mapping from form cells to realized cells, as shown in Table 5. The former possibility is exemplified by cases where overabundance is due to the availability of different suppletive stems – for instance, *muls-* and *mulct-* as the stem on which the supine forms of the Latin verb meaning 'milk' are built – or phonologically identical stems belonging to different inflection classes – for instance, the Latin verb meaning 'clean' can have its imperfective forms inflected according to either the 2nd or the 3rd conjugation – as form correspondents of the same content cell. The latter is the way in which cases of completely systematic overabundance across the whole lexicon arise. For instance, in Latin two different endings, *-rīs* and *-re*, are available for the second-person singular of the passive across all tenses and moods, as exemplified by the present indicative in Table 5. In this case, the mapping from content cells to form cells is straightforward, as there is a single stem univocally assigned to a specific inflection class; overabundance arises in the mapping from form cells to realized cells, because there are two realization rules that can be applied to the form cell at hand, one introducing *-rīs*, the other introducing *-re*.

Going back to the topic of this section, perhaps it is the way in which overabundance arises, rather than the sheer number of lexemes involved, that is related to the usefulness of a treatment introducing distinct flexemes:[7] this might be considered as more appropriate to model overabundance phenomena that arise in the mapping from content cells to form cells because of the availability of different suppletive stems or phonologically identical stems that belong to different inflection classes, and less

[7] The two aspects do not go necessarily hand in hand: for instance, the overabundance phenomena discussed in Sect. 4.3 below – i.e., overabundance in perfective cells of 4th-conjugation verbs in Latin – can be considered as being introduced by a realization rule that, however, is applied only in a specific inflection class, rather than across the whole lexicon.

Table 6 A portion of the paradigm of PONO 'place' in the tabular format

content	form
PRS.ACT.IND.1SG	*ponō*
PRS.PASS.IND.2SG	*ponerīs,ponere*
PRF.ACT.IND.1SG	*posuī,posīvī*
FUT.ACT.IND.1SG	*ponam*
PRS.ACT.SBJV.1SG	*ponam*
FUT.ACT.IMP.2SG	*ponitō*
FUT.ACT.IMP.3SG	*ponitō*

appropriate to model overabundance phenomena that arise in the mapping from form cells to realized cells because of the availability of different realization rules. Indeed, also the cases of overabundance that Bonami and Crysmann (2018) discuss to argue for the use of flexemes in a formal theory of grammar can ultimately be reconducted to the former type.[8]

Yet another way to account for cases of completely systematic overabundance in a given cell in all lexemes would be to treat the two variants as belonging to separate cells, in a purely formal conception of the shape of paradigms. A short digression is needed to bring some support to this option. Boyé and Schalchli (2016) have shown that different representations of a lexeme's paradigm are conceivable, each with its own strengths and weaknesses. The "tabular paradigm" is the format that is most usual in language descriptions, and it is driven by content: the different morphosyntactic property sets for which the lexemes of the lexical category under consideration can be inflected are the starting point, and for each of them the corresponding wordform is given, as shown in Table 6 on a portion of the paradigm of the Latin verb PONO 'place', chosen because it displays non-canonical phenomena like syncretism (of FUT.ACT.IND.1SG with PRS.ACT.SBJV.1SG and of FUT.ACT.IMP.2SG with FUT.ACT.IMP.3SG) and overabundance (in PRS.PASS.IND.2SG and in PRF.ACT.IND.1SG) with different degrees of generality across the lexicon.

Alternatively, it is possible to start from the different wordforms available for a given lexeme, and map each of them to the morphosyntactic property sets that they can express: this corresponds to what Boyé and Schalchli (2016: 209) call the "morphomic paradigm", which is purely driven by form, and is illustrated with the same Latin data in Table 7. A disadvantage of this representation format is that paradigm size is not uniform, as patterns of syncretism and overabundance can be variable across lexemes.

As a more balanced format, Boyé and Schalchli (2016: 209 ff.) propose to use what they call the "cell paradigm". This can be obtained by starting from the content-driven tabular paradigm, where different contents belong to different cells, and abstracting away from cases of syncretism (as in the morphomic paradigm), but only if

[8] However, it should be observed that the distinction between overabundance arising in the mapping from content cells to form cells and overabundance arising in the mapping from form cells to realized cells is not always straightforward, especially when it is caused by the presence of different exponents, that could be analyzed either as deriving from different inflection class assignments or as being introduced by different realization rules.

Table 7 A portion of the paradigm of PONO 'place' in the morphomic format

form	content
ponō	PRS.ACT.IND.1SG
ponerīs	PRS.PASS.IND.2SG
ponere	PRS.PASS.IND.2SG
posuī	PRF.ACT.IND.1SG
posīvī	PRF.ACT.IND.1SG
ponam	FUT.ACT.IND.1SG,PRS.ACT.SBJV.1SG
ponitō	FUT.ACT.IMP.2SG,FUT.ACT.IMP.3SG

they are completely systematic across lexemes, thus keeping paradigm size uniform within lexical categories. Therefore, FUT.ACT.IMP.2SG and FUT.ACT.IMP.3SG are conflated in a single cell, because there is syncretism between these two morphosyntactic property sets in all Latin verbs, while FUT.ACT.IND.1SG and PRS.ACT.SBJV.1SG are kept separate, because there is no syncretism between these two cells in 1st and 2nd-conjugation verbs (cf. AMO 'love' with FUT.ACT.IND.1SG *amābō* and PRS.ACT.SBJV.1SG *amem*). The outcome of this procedure looks like in Table 8, where it can be observed that cases of overabundance are treated in the same way as in the tabular paradigm, regardless of their systematicity. However, if we want to focus on the formal, rather than semantic, structure of paradigms, it seems reasonable to abstract away not only from cases of completely systematic syncretism, as suggested by Boyé and Schalchli (2016), but also from cases of completely systematic overabundance. In this way, there would be two cells that correspond to the morphosyntactic property set PRS.PASS.IND.2SG, one for the form in *-rīs* and the other for the form in *-re*: since these two variants are available for all Latin lexemes, it makes sense to say that they fill two different slots, at least from a purely formal perspective. On the other hand, *posuī* and *posīvī* still fill a single cell, because not all Latin verbs are overabundant in PRF.ACT.IND.1SG. This representation format is illustrated in Table 9.

Arguably, different representation formats have different (dis)advantages, and the preference for one rather than another ultimately depends on the purpose. Going back to the topic of the present paper, the representation format of the cell paradigm as exemplified in Table 9 would have the advantage of providing an alternative strategy for exactly the cases in which the introduction of distinct flexemes intuitively seems less satisfying as a way to model overabundance – i.e., the most systematic ones that are caused by the availability of different realization rules in the mapping from form cells to realized cells (cf. Table 5 above).

2.5 Interim summary

To sum up, we have seen that some of the dimensions of variation identified in Thornton's (2019) canonical typology of overabundance seem to be orthogonal to the issue of the usefulness of flexemes as a modelling strategy: this is the case of the criterion based on the presence (or absence) of conditions, and of the one based on the strength of overabundance as measured by the frequency ratio between the cell mates.

Table 8 A portion of the paradigm of PONO 'place' in the cell format, as defined by Boyé and Schalchli (2016)

content	form
PRS.ACT.IND.1SG	*ponō*
PRS.PASS.IND.2SG	*ponerīs,ponere*
PRF.ACT.IND.1SG	*posuī,posīvī*
FUT.ACT.IND.1SG	*ponam*
PRS.ACT.SBJV.1SG	*ponam*
FUT.ACT.IMP.2SG,FUT.ACT.IMP.3SG	*ponitō*

Table 9 A portion of the paradigm of PONO 'place' in the cell format, in our proposal

content	form
PRS.ACT.IND.1SG	*ponō*
PRS.PASS.IND.2SG (A)	*ponerīs*
PRS.PASS.IND.2SG (B)	*ponere*
PRF.ACT.IND.1SG	*posuī,posīvī*
FUT.ACT.IND.1SG	*ponam*
PRS.ACT.SBJV.1SG	*ponam*
FUT.ACT.IMP.2SG,FUT.ACT.IMP.3SG	*ponito*

We have then discussed potential interactions of different degrees of systematicity of overabundance across cells and/or lexemes on the one hand, and the appeal of a modelling using flexemes on the other hand, showing that alternative options are available for overabundance phenomena that involve a single cell or lexeme, but also for overabundance phenomena that involve all lexemes.

As for the former case, we argue that if overabundance is modelled by means of flexemes, then they should be used also for the most idiosyncratic cases where only one cell or lexeme is involved. Indeed, flexemes can be defined as lexical items characterized by a unique set of inflected wordforms, as it appears to follow both from the definitions proposed by the scholars that introduced them in the first place (cf. Fradin & Kerleroux, 2003: 193, Fradin, 2019: 71) and from the way they have been applied to overabundance in subsequent work by Thornton (2018) and Bonami and Crysmann (2018). If that is true, then it seems contradictory to have overabundant flexemes, even if this happens in a single cell and/or lexeme, as this would mean that their set of inflected wordforms is not unique. Furthermore, if the appeal of flexemes consists exactly in the fact that they can capture the systematicity of overabundance, either across cells or across lexemes (as it emerges from the discussion above), then even in cases of overabundance in a single cell, flexemes may still be useful to capture systematicity across lexemes, and, conversely, in cases of overabundance in a single lexeme, they may still be useful to capture systematicity across cells. For instance, the use of flexemes to model overabundance in a single cell – the plural – of Italian nouns such as LENZUOLO allows us to capture the fact that the inflectional behavior of LENZUOLO$_1$ and LENZUOLO$_2$ is comparable to the one of non-overabundant nouns that belong to the same inflection classes, like LIBRO 'book' and UOVO 'egg', respectively, as shown in Table 10. Conversely, using flexemes to account for the

Table 10 Flexemes capturing generalizations across lexemes

lexeme	flexeme	SG	PL	Inflection class
LENZUOLO	LENZUOLO$_1$	*lenzuolo*	*lenzuoli*	1 (*-o*/*-i*)
	LENZUOLO$_2$	*lenzuolo*	*lenzuola*	5 (*-o*/*-a*)
LIBRO	LIBRO	*libro*	*libri*	1 (*-o*/*-i*)
UOVO	UOVO	*uovo*	*uova*	5 (*-o*/*-a*)

Table 11 Flexemes capturing generalizations across cells

lexeme	flexeme	PRS.IND					
		1SG	2SG	3SG	1PL	2PL	3PL
DOVERE	DOVERE$_1$	*devo*	*devi*	*deve*	*dobbiamo*	*dovete*	*devono*
	DOVERE$_2$	*debbo*	*devi*	*deve*	*dobbiamo*	*dovete*	*debbono*

idiosyncratic overabundance phenomenon found only in the Italian verb DOVERE ‘must’ allows us to capture the relationship between all variants with stem *dev-* on the one hand, and all variants with stem *debb-* on the other hand, across the different cells in which they are found, as shown in Table 11.

Only for cases of overabundance in all lexemes in a given cell we propose a different treatment – i.e., splitting the cell in two, as suggested in Sect. 2.4. This is motivated by the lack of economy of the introduction of distinct flexemes for all lexemes that would be required by such cases. Furthermore, the use of flexemes can ultimately be considered as implying a purely formal approach to the identification of lexical items: therefore, it is fully consistent with the adoption of a purely form-driven representation format for paradigms, like the one that is implied by this solution, that would have the obvious advantage of avoiding having to map each lexeme to different flexemes.

3 From theory to practice: building a flexicon of Latin principal parts

To understand the relevance of the theoretical questions addressed in this paper, it is useful to spend a few words on the practical objectives that led us to investigate them, in the context of resource development for the Latin language.

The LiLa project (Passarotti et al., 2020) aims at creating a Knowledge Base of interoperable linguistic resources for the Latin language, following the principles of the (Linguistic) Linked (Open) Data paradigm (Cimiano et al., 2020). This means using Semantic Web standards like RDF and SPARQL to represent and query data, respectively, and sharing a common vocabulary for knowledge representation by reusing existing ontologies – the most relevant being OLiA (Chiarcos & Sukhareva, 2015) for linguistic annotation, NIF (Hellmann et al., 2013) and CoNLL-RDF (Chiarcos & Fäth, 2017) for corpus annotation and OntoLex-Lemon (Buitelaar et al., 2011; McCrae et al., 2017) for lexical information – and building new ones if needed – for

Table 12 Conjugation of Latin verbs

verb (meaning)	conj.	PRS.ACT.INF	PRF.ACT.IND.1SG	SUP.ACC
AMO (‘love’)	1st	*amāre*	*amāvī*	*amātum*
MONEO (‘warn’)	2nd	*monēre*	*monuī*	*monitum*
LEGO (‘read’)	3rd	*legere*	*lēgī*	*lectum*
VENIO (‘come’)	4th	*venīre*	*vēnī*	*ventum*
CAPIO (‘take’)	mixed[9]	*capere*	*cepī*	*captum*

Table 13 Declensions of Latin nouns

noun (meaning)	decl.	GEN.SG
ROSA (‘rose’)	1st	*rosae*
LUPUS (‘wolf’)	2nd	*lupī*
ANIMAL (‘animal’)	3rd	*animālis*
MANUS (‘hand’)	4th	*manūs*
RES (‘thing’)	5th	*reī*

Table 14 Classes of Latin adjectives

adjective (meaning)	class	NOM.F.SG	NOM.N.SG
BONUS (‘good’)	1st	*bona*	*bonum*
ACER (‘sharp’)	2nd	*acris*	*acre*

instance, the LiLa ontology itself and the one designed for the inclusion of the Word Formation Latin derivational lexicon (cf. Litta & Passarotti, 2019; Pellegrini et al., 2021). Structurally, at the core of the architecture of the LiLa Knowledge Base, there is the so-called Lemma Bank, a collection of citation forms of Latin words. Interoperability between the different resources linked to the Knowledge Base is achieved by connecting the entries of lexical resources and the tokens of textual resources to their respective citation form in the Lemma Bank.

Currently, very little information on inflectional morphology is available for the lemmas represented in the Knowledge Base. In the Lemma Bank, only coarse-grained information on their traditional conjugation (for verbs), declension (for nouns) or class (for adjectives) is provided alongside each citation form. Table 12, Table 13 and Table 14 illustrate these traditional classifications and the inflectional behavior of lexemes of different classes in a few exemplary cells.

However, these classifications are not sufficient to capture the inflectional behavior of Latin lexemes in its entirety, as they neglect the role played by stem allomorphy. For instance, the verb DELEO (‘delete’) is traditionally assigned to the 2nd conjugation like MONEO in Table 12, because they share the same endings in wordforms based on the present stem, like PRS.ACT.INF (*delēre* like *monēre*), but the two verbs differ in the way they form their PRF.ACT.IND.1SG, based on the perfect stem (*delēv-ī*

Table 15 Overabundance in the principal parts of AC(C)OM(M)ODO 'adjust'

lexeme	PRS.ACT.INF	PRF.ACT.IND.1SG	SUP.ACC
AC(C)OM(M)ODO	*accommodāre*,*accomodāre*, *acommodāre*,*acomodāre*	*accommodāvī*,*accomodāvī*, *acommodāvī*,*acomodāvī*	*accommodātum*,*accomodātum*, *acommodātum*,*acomodātum*

vs. *monu-ī*) and SUP.ACC, based on what Aronoff (1994: 35) calls the third stem (*delēt-um* vs. *monit-um*).

Indeed, alongside the citation form of each entry, usually Latin dictionaries also list a set of principal parts from which the other inflected wordforms can be inferred (cf. Stump & Finkel, 2013). The wordforms that are used as principal parts in the Oxford Latin Dictionary (Glare, 2012) are the ones that we report in Table 12, Table 13 and Table 14. Although at least one dictionary (namely, Lewis & Short, 1879) has been made available as a digital resource linked to the LiLa Knowledge Base (cf. Mambrini et al., 2021), the focus was on semantic information, rather than on form, so principal parts are not provided in that place either. This means that principal parts are not represented in the Lemma Bank, nor in other resources linked to the Knowledge Base.[10]

To fill this gap, we are working at the creation and linking to the Lemma Bank of PrinParLat,[11] a resource listing the principal parts of Latin verbs, nouns and adjectives and providing a fine-grained inflectional classification that would allow users to generate full paradigms if needed. Since the bulk of the Lemma Bank has been obtained from the database of Lemlat 3.0, a recently renewed morphological analyzer (cf. Passarotti et al., 2017), and also information on stems and inflection classes is available in Lemlat's database, it is reasonable to exploit such information to obtain principal parts. However, as Lemlat is designed to analyze wordforms, rather than to produce them, for some lexemes many different variants are represented. For instance, for the verb meaning 'adjust', Lemlat lists four different variants of the present stem, namely, *accommod-*, *accomod-*, *acommod-* and *acomod-*. Therefore, four variants can be generated for each of the principal parts, as shown in Table 15.

To account for this state of affairs, our idea is to create a flexicon, rather than a lexicon, that is, a resource listing flexemes, rather than lexemes, following the hint of Walther (2013). Since the resource at hand only provides formal information, it is reasonable to rely on formal units (i.e., flexemes) rather than on semantic ones

[9] We follow Dressler (2002: 107) and use the label of "mixed conjugation" because the members of this class are heteroclite lexemes that inflect like 3rd-conjugation verbs in some cells (e.g., in PRS.ACT.INF), and like 4th-conjugation verbs in other cells (e.g., in the citation form).

[10] The only exceptions are the past participles of verbs, that are sometimes used instead of SUP.ACC as a principal part allowing to infer wordforms built on the third stem and are represented in the Lemma Bank as instances of so-called "hypolemmas". This is motivated by the fact that some resources may lemmatize participial forms of a verb like PRF.PTCP.NOM.M.SG *amātus* under the corresponding adjective rather than under the verb itself (in this case, under AMATUS rather than AMO). Hence, also the participial form should be available as a special kind of lemma linked to its overarching main verbal lemma. The reader is referred to Passarotti et al. (2020: 190 ff.) for further technical details.

[11] https://lila-erc.eu/data/lexicalResources/prinparlat/Lexicon. At the time of the publication of this paper, only verbs are included in the resource.

Table 16 The four flexemes of the lexeme AC(C)OM(M)ODO ‘adjust’

lexeme	flexeme	PRS.ACT.INF	PRF.ACT.IND.1SG	SUP.ACC
AC(C)OM(M)ODO	ACCOMMODO	*accommodāre*	*accommodāvī*	*accommodātum*
	ACCOMODO	*accomodāre*	*accomodāvī*	*accomodātum*
	ACOMMODO	*acommodāre*	*acommodāvī*	*acommodātum*
	ACOMODO	*acomodāre*	*acomodāvī*	*acomodātum*

(i.e., lexemes). More importantly, introducing a different flexeme for each stem variant (as illustrated in Table 16) seems to be a promising way to be able to generate all the corresponding wordforms, rather than selecting only one per cell as in other similar resources (cf. Pellegrini & Passarotti, 2018), and at the same time to capture the systematic relation that holds between the wordforms that are built on the same stem variant in different cells, rather than just listing them as an unstructured set of alternatives, potentially independent of each other, as in the format of Table 15.

However, this requires us to have a method to extract the relevant units from the large-scale data we are dealing with in a principled fashion: this is the topic of the next section.

4 Identifying systematicity

In this section, we describe how a modelling of overabundance by means of flexemes is implemented in the context of the resource of Latin principal parts mentioned in the previous section. We are going to focus on verb inflection, where the picture is more complex for what concerns patterns of allomorphy – besides the five conjugations mentioned above, there is widespread stem allomorphy in the relation between the present stem, the perfect stem and the third stem, with wordforms based on different stems being not fully predictable from each other – and, consequently, overabundance in different cells. In Sect. 4.1, we detail the way in which the information given in Lemlat’s database has been exploited to generate principal parts. In Sect. 4.2, we propose an abstractive method to organize the cell mates in different flexemes in a principled fashion. We then evaluate the outcome of the proposed method qualitatively in Sect. 4.3, where we also discuss some cases that are potentially problematic, and quantitatively in Sect. 4.4, where we show how much we can reduce the number of possible flexemes.

4.1 The data

The database of Lemlat is organized around a collection of LESs (“LExical Segments”), roughly corresponding to stems used in different portions of the paradigm. To each LES, a code (called CODLES) is assigned that gives information on which wordforms can be built on that LES, and which endings are compatible with it. For instance, the LESs and CODLESs reported in Lemlat’s database for BENEDICO ‘praise’ are shown in the two columns on the left in Table 17.

Table 17 Generation of principal parts of BENEDICO 'praise'

information in Lemlat		generated principal parts	
LES	CODLES	cell	form
benedic	v3r	PRS.ACT.INF FUT.ACT.IND.3SG	*benedic-ere* *benedic-et*
benedix	v7s	PRF.ACT.IND.1SG	*benedix-ī*
benedict	n41	/	/
benedict	n6p1	PRF.PTCP.NOM.N.SG	*benedict-um*
benedictur	n6p2	FUT.PTCP.NOM.N.SG	*benedictur-um*

The CODLES "v3r" tells us that the LES *benedic-* can be used to generate wordforms belonging to the portion of paradigm that is traditionally called the present system – i.e., imperfective wordforms – by appending endings of the 3rd conjugation to it; the CODLES "v7s" tells us that the LES *benedix-* can be used to generate wordforms of the so-called perfect system – i.e., perfective wordforms – by appending the regular endings that are used across all conjugations; the CODLESs "n41", "n6p1" and "n6p2" identify the corresponding LESs as the ones on which wordforms of the supine, perfect participle and future participle can be formed, respectively. These are all wordforms that are considered as being based on the third stem in Aronoff's (1994) analysis.

There are two reasons why Lemlat makes use of three different LESs that all correspond to Aronoff's third stem. The first one is that in Lemlat, some verbs are considered to lack one of these stems, but not others: for instance, they may lack the perfect participle, but not the supine, as happens for VENIO 'come'. In such cases, the presence of a LES with a given CODLES indicates that wordforms built on that LES may be encountered in texts, its absence that this is not possible, hence the corresponding wordforms should not be generated. The second one is that the stem on which the future participle is built is not always predictable on the basis of the one on which the perfect participle and the supine are built, as happens in most cases, where the former can be obtained by adding *-ur-* to the latter: there are a handful of verbs that display an unpredictable allomorph in the future participle, for instance MORIOR 'die', with PRF.PTCP.NOM.N.SG *mortu-um* but FUT.PTCP.NOM.N.SG *morit-ur-um*. This makes it necessary to have an additional slot in which this information can be coded, and hence a separate CODLES for the future participle.

All this information can be exploited to generate principal parts, as shown in the two columns on the right in Table 17. As recognized in traditional descriptions and confirmed by recent theoretical studies (Finkel & Stump, 2009; Pellegrini, 2020), two principal parts are needed to account for the wordforms of the Present System. This is due to the presence of a quite large class of lexemes that are heteroclite between the 3rd and 4th conjugation, those of the so-called mixed class: to be able to identify their class, it is necessary to have one principal part that is inflected according to the 3rd conjugation (from which all the other 3rd-conjugation wordforms can be inferred) and one principal part that is inflected according to the 4th conjugation (from which all the other 4th-conjugation wordforms can be inferred). Most Latin grammars and dictionaries use PRS.ACT.INF for the former and the citation form, PRS.ACT.IND.1SG, for the latter. In our resource, we follow the tradition in using PRS.ACT.INF, but we depart from it in discarding PRS.ACT.IND.1SG. This is motivated by the fact that

Table 18 Generation of principal parts of FUGO 'cause to flee'

information in Lemlat		generated principal parts	
LES	CODLES	cell	form
fug	v1r	PRS.ACT.INF	*fug-āre*
		FUT.ACT.IND.3SG	*fug-ābit*
		PRF.ACT.IND.1SG	*fug-avī*
		PRF.PTCP.NOM.N.SG	*fug-atum*
		FUT.PTCP.NOM.N.SG	*fug-atūrum*

PRS.ACT.IND.1SG is poorly informative on the content of other cells, since in this cell the distinction between the 1st and 3rd conjugation is neutralized. Furthermore, being the citation form, it is already reported in the Lemma Bank of the LiLa Knowledge Base, to which the resource will be linked, so having it in the resource too would be redundant. Therefore, the principal part that we select to account for other wordforms of the present system is FUT.ACT.IND.3SG, that is more informative on the conjugation of the corresponding verb. For deponent verbs, that are defective of morphologically active wordforms, we use the corresponding morphologically passive cells as principal parts, i.e., PRS.PASS.INF and FUT.PASS.IND.3SG, for instance *potīrī* and *potiētur* for POTIOR 'obtain'.

As for the Perfect System, a single principal part is sufficient, and we follow the traditional usage in selecting PRF.ACT.IND.1SG. The passive cells of the perfect system do not have a dedicated synthetic form but are filled with a periphrase composed of the perfect participle and the appropriately inflected form of the verb 'be' (e.g., PRF.ACT.IND.1SG *amātus sum* 'I was loved'). As a consequence, for deponents all the cells of the perfect system are filled periphrastically, hence we cannot provide a synthetic wordform as principal part for these cells. For impersonal verbs, like DECET 'be suitable', we use PRF.ACT.IND.3SG (*decuit*) instead of PRF.ACT.IND.1SG, that is lacking.

As for the portion of the paradigm that Aronoff (1994) considers as being based on the third stem, we provide two principal parts, PRF.PTCP.NOM.N.SG and FUT.PTCP.NOM.N.SG, to account for the abovementioned cases where the perfect and future participle display different allomorphs. We do not give a principal part for the supine because this information is not systematically provided for all verbs in Lemlat's database. Note that for verbs that are marked as fully regular in Lemlat's database, all the relevant principal parts can be generated from the single LES that is given, i.e., the one that corresponds to the present stem, as shown in Table 18.

A last remark that should be made on the principal parts that we generate concerns vowel length. For the purposes of this paper, this information is systematically coded in the endings that we append, at least in those cases in which we always know whether the vowels are long or short. However, this information is currently lacking in our data for what concerns the stems, as it is not given in Lemlat's database, and it is not straightforward to systematically obtain it from other sources. In the future,

Table 19 Generation of principal parts and flexemes of MUNERO/MOENERO 'bestow'

information in Lemlat		generated principal parts		
LES	CODLES	cell	form	flexeme
muner	v1r	PRS.ACT.INF	*muner-āre*	MUNERO
		FUT.ACT.IND.3SG	*muner-ābit*	MUNERO
		PRF.ACT.IND.1SG	*muner-avī*	MUNERO
		PRF.PTCP.NOM.N.SG	*muner-atum*	MUNERO
		FUT.PTCP.NOM.N.SG	*muner-aturum*	MUNERO
moener	v1r	PRS.ACT.INF	*moener-āre*	MOENERO
		FUT.ACT.IND.3SG	*moener-ābit*	MOENERO
		PRF.ACT.IND.1SG	*moener-avī*	MOENERO
		PRF.PTCP.NOM.N.SG	*moener-atum*	MOENERO
		FUT.PTCP.NOM.N.SG	*moener-atūrum*	MOENERO

Table 20 Generation of principal parts and flexemes of CLU(E)O 'be named'

information in Lemlat		generated principal parts		
LES	CODLES	cell	form	flexeme
clu	v2r	PRS.ACT.INF	*clu-ēre*	CLUEO
		FUT.ACT.IND.3SG	*clu-ēbit*	CLUEO
clu	v3r	PRS.ACT.INF	*clu-ere*	CLUO
		FUT.ACT.IND.3SG	*clu-et*	CLUO

we plan to extract this information and incorporate it in our resource,[12] as in some cases the difference between stems – and consequently between wordforms based on them – is exactly vowel length: see the contrast between PRS.ACT.IND.3SG *ven-it* and PRF.ACT.IND.3SG *vēn-it* in the verb VENIO 'come'. However, it should be borne in mind that the results that we show in the reminder of this paper do not take this aspect into consideration. This is reflected by the spellings that are adopted from now on, that are the ones that were actually used for the purposes of this paper rather than the ones normally found in Latin grammars and dictionaries.

In this picture, overabundance emerges when more than one LES is available from which different wordforms that fill the same cell can be generated, as in the data reported in Table 19, Table 20, Table 21 and Table 22. In the discussion of Sect. 2.4, it was suggested that flexemes might be more useful when overabundance arises because of the availability of suppletive stem variants, or phonologically identical stems that are assigned to different inflection classes. Therefore, one might be tempted to adopt a constructive approach to the identification of flexemes, introducing distinct flexemes only when for a given principal part, Lemlat's database reports two (or more) segmentally different LESs – i.e., stems – or assign different CODLESs – i.e., inflection classes – to segmentally identical LESs. Two cell mates would then be assigned to the same flexeme only if they are built on the same stem (cf. Table 19) or according to the same inflection class (cf. Table 20)

However, things are not always so simple, and in some cases such a method would run into serious problems. A general problem is that the Latin inflection class system is "segregated" in the terms of Stump (2016: 90), meaning that the traditional conju-

[12] In the version released at the time of the publication of this paper, information on vowel length is not coded at all – not even on the endings – to avoid having a notation that only concerns a portion of forms, as this might be confusing for users.

Table 21 Generation of principal parts and flexemes of EXTERG(E)O ‘wipe clean’

information in Lemlat		generated principal parts		
LES	CODLES	cell	form	flexeme
exterg	v2r	PRS.ACT.INF	*exterg-ēre*	EXTERGEO
		FUT.ACT.IND.3SG	*exterg-ēbit*	EXTERGEO
exterg	v3r	PRS.ACT.INF	*exterg-ere*	EXTERGO
		FUT.ACT.IND.3SG	*exterg-et*	EXTERGO
exters	v7s	PRF.ACT.IND.1SG	*exters-ī*	?
exters	n6p1	PRF.PTCP.NOM.N.SG	*exters-um*	?
extersur	n6p2	FUT.PTCP.NOM.N.SG	*extersur-um*	?

gations are only strictly relevant in the present system. This raises the question of how to handle the relationship with wordforms outside the present system. Let us consider, for instance, the data of Table 21. In the present system, the situation is exactly like in Table 20: the LES *exterg-* can be assigned either to the 2nd or the 3rd conjugation, so two variants of the principal parts PRS.ACT.INF and FUT.ACT.IND.3SG can be generated accordingly. However, while for CLUEO no other LES is provided by Lemlat, and consequently no other principal part can be generated, for EXTERG(E)O Lemlat also gives additional LESs corresponding to the stems on which perfective and participial wordforms are built, so we can generate the corresponding principal parts. The problem is deciding the flexeme that should be assigned to these additional, non-overabundant principal parts: is it the one of 2nd-conjugation wordforms, the one of 3rd-conjugation wordforms, both of them, or neither of them? It is not easy to give a principled answer to this question, but in any event the conjugation of the verb is uninformative in this respect.

The data summarized in Table 22 raise other interesting problems. Again, the assignment of flexemes is straightforward in the present system, but it is more difficult to identify the relation with other principal parts, and the relation of these other principal parts with one another. Intuitively, we might want to identify a flexeme including all the variants with the nasal segment *-n-* (*pangere, panget, panxi, panctus*), but this cannot be captured simply by stem identity. As for the other perfective and participial wordforms, there is not even a clear intuition as to the flexeme to which they should be assigned. In this case, too, the method outlined above does not help us to take a principled decision.

From these examples, we conclude that relying only on the fact that two cell mates are built on the same stem or belong to the same inflection class is not a satisfying way of assigning them to flexemes. This appears to be ultimately due to the fact that this method is based on a constructive perspective – in the terms of Blevins (2016) – where wordforms are considered to be built from stems by means of realization rules sensible to the inflection class to which those stems are assigned. However, it has been shown that segmenting forms in a stem and one or more exponents in a systematic fashion is a difficult task, not only in Latin, as proven by the data discussed in this section, but in many languages with complex inflectional morphology (cf. Beniamine, 2018: 74 ff. and Bonami & Beniamine, 2021 for relevant discussion). This is what motivates our attempt at an abstractive method to identify cell mates that should be assigned to the same flexeme, using only information that can be inferred from unsegmented wordforms. In the next subsection, we will discuss our method in detail.

Table 22 Generation of principal parts and flexemes of PANGO/PAGO/PACO ‘fasten’

information in Lemlat		generated principal parts		
LES	CODLES	cell	form	flexeme
pang	v3r	PRS.ACT.INF	*pang-ere*	PANGO
		FUT.ACT.IND.3SG	*pang-et*	PANGO
pag	v3r	PRS.ACT.INF	*pag-ere*	PAGO
		FUT.ACT.IND.3SG	*pag-et*	PAGO
pac	v3r	PRS.ACT.INF	*pac-ere*	PACO
		FUT.ACT.IND.3SG	*pac-et*	PACO
panx	v7s	PRF.ACT.IND.1SG	*panx-ī*	?
peg	v7s	PRF.ACT.IND.1SG	*peg-ī*	?
pepig	v7s	PRF.ACT.IND.1SG	*pepig-ī*	?
panct	n6p1	PRF.PTCP.NOM.N.SG	*panct-um*	?
pact	n6p1	PRF.PTCP.NOM.N.SG	*pact-um*	?

4.2 The method

We start from two related – and strong – assumptions on the application of flexemes to cases of overabundance. Namely, we assume that i) flexemes cannot be overabundant, i.e., each flexeme cannot have more than one wordform for each paradigm cell, as we have motivated in Sect. 2.5; and, consequently, that ii) distinct flexemes should be introduced to account for any case of overabundance.[13]

To illustrate our procedure to identify flexemes, it is useful to enucleate what motivates the intuition that we have in clear-cut cases. For instance, if we go back at the data of Table 20, the intuition that *cluēre* and *cluēbit* should be assigned to the same flexeme is justified by the fact that in Latin there are many verbs that have a PRS.ACT.INF in *-ēre* and a FUT.ACT.IND.3SG in *ēbit*: these are all the verbs that are traditionally assigned to the 2nd conjugation. Hence, it is reasonable to consider these two wordforms as patterning together in the same way as those of non-overabundant 2nd-conjugation verbs. Similarly, there are many verbs – those of the 3rd conjugation – that have a PRS.ACT.INF in *-ere* and a FUT.ACT.IND.3SG in *-et*, so it makes sense to assign *cluere* and *cluet* to the same flexeme. Conversely, no Latin verb (except for those that display a pattern of overabundance similar to the one of CLU(E)O) has a PRS.ACT.INF in *-ēre* and a FUT.ACT.IND.3SG in *-et* or a PRS.ACT.INF in *-ere* and a FUT.ACT.IND.3SG in *-ēbit*, bringing further support to the intuition that *cluēre* and *cluet* should be assigned to different flexemes, as do *cluere* and *cluēbit*. The method that we propose builds on this intuition, applying it to all possible pairs of flexemes in a purely word-based, abstractive fashion.

Concretely, we start from all the possible combinations of all the principal parts that can be generated for a given lexeme in different cells. Let us consider again the case of CLU(E)O, where we have two cell mates for each of the two principal parts that we can generate from the information provided by Lemlat. For some verbs, some principal parts are not generated, and the corresponding cells are left empty, because the relevant LESs are not reported in Lemlat’s database: this is the case of the PRF.ACT.IND.1SG, PRF.PTCP.NOM.N.SG and FUT.PTCP.NOM.N.SG of CLU(E)O.

[13] Except for the ones that are completely systematic across lexemes and can thus be treated by splitting the cell(s) at hand in two, as suggested in Sect. 2.5.

Table 23 Outcomes available for the principal parts of CLU(E)O

PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
cluēre;cluere;/	*cluēbit;cluet;/*	/	/	/

Table 24 Possible combinations of the outcomes available for the principal parts of CLU(E)O

	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
1	***cluēre***	***cluēbit***	/	/	/
2	*cluēre*	*cluet*	/	/	/
3	*cluēre*	/	/	/	/
4	*cluere*	*cluēbit*	/	/	/
5	***cluere***	***cluet***	/	/	/
6	*cluere*	/	/	/	/
7	/	*cluēbit*	/	/	/
8	/	*cluet*	/	/	/

Therefore, we had to deal also with combinations involving empty cells. To be able to make the method work, we need to consider as a valid flexeme also one that only consists of a combination of wordforms of a subset of cells. Technically, this amounts at considering also being empty as an additional possible value of any cell, both in cases where an actual wordform is attested and in cases where it is not, as shown in Table 23, where the empty cells are represented by a slash.

Therefore, we have three possibilities in PRS.ACT.INF and FUT.ACT.IND.3SG (the two available cell mates and the empty cell represented by the slash), and the empty cell as the only possibility for the other principal parts. This gives us $3 \times 3 \times 1 \times 1 \times 1 - 1 = 8$ possible combinations of the different possibilities, as shown in Table 24.

On the basis of the considerations made in the discussion of Sect. 2.5 above, each of these combinations corresponds to a possible flexeme, as it is a unique set of wordforms with no overabundance. What we want to do is selecting the flexemes that should be kept as valid among them. To this aim, the first step of our procedure consists in looking at the formal relation between each of the combinations of the different possibilities available for all cells, pair by pair. For each pair of cells, we extract alternation patterns between each pair of wordforms using the Qumin toolkit (Beniamine, 2018). The alternation pattern is computed by simply comparing two wordforms and isolating the segments that change from one wordform to the other one, abstracting away from the ones that are shared by the two wordforms (cf. the third column of Table 25).[14] We then use these patterns to decide which of the combinations should be kept as valid flexemes, and which ones should be dis-

[14] More precisely, we used the find_patterns.py script, selecting the patternsPhonsim option, that relies on an algorithm that exploits information on the phonological similarity of the wordforms. For this purpose, wordforms had to be converted to a (not fully informative) phonological transcription, based on the one adopted in Pellegrini (2020). For more details on the identification of patterns, the reader is referred to Beniamine (2018: 88 ff.) and to the web documentation of the toolkit (https://qumin.readthedocs.io/).

Table 25 Possible combinations of the outcomes available for the PRS.ACT.INF and FUT.ACT.IND.3SG of CLU(E)O

	PRS.ACT.INF	FUT.ACT.IND.3SG	alternation pattern	non-overabundant verbs with the same alternation pattern
1	***cluēre***	***cluēbit***	**re ⇌ bit**	**HABEO, MONEO, MULGEO ...** **(all 2nd-conjugation verbs)**
2	*cluēre*	*cluet*	ēre ⇌ et	no verb
3	***cluēre***	**/**	**/**	
4	*cluere*	*cluēbit*	ere ⇌ ēbit	no verb
5	***cluere***	***cluet***	**re ⇌ t**	**CADO, DICO, LEGO, ...** **(all 3rd-conjugation verbs)**
6	***cluere***	**/**	**/**	
7	**/**	***cluēbit***	**/**	
8	**/**	***cluet***	**/**	

carded. Table 25 shows the way in which the choice is made for the PRS.ACT.INF and FUT.ACT.IND.3SG of CLU(E)O, highlighting in bold the combinations that are kept, namely i) the combinations such that the alternation pattern that holds between the wordforms is attested in at least one verb that does not display overabundance in either of the cells at hand (the ones in lines 1 and 5), and ii) the combinations involving one empty cell (the ones in lines 3, 6, 7 and 8). The combinations in lines 2 and 4 are discarded because their alternation pattern is not attested in verbs that are not overabundant in these cells.

The same operation is performed on all the other combinations of cells of the verb under consideration. In this case, since in the other principal parts the only available possibility is the empty cell, all other combinations would be considered as valid.

The next step consists in scaling up from combinations of pairs of cells to combinations of all possible cells – the ones of Table 24. A combination is kept as a flexeme if all the wordforms pattern together according to the pairwise comparison described above. Lastly, flexemes that are a subset of another flexeme are removed. This allows us to exclude the flexeme in line 3 of Table 24, with *cluēre* in PRS.ACT.INF and no wordform in all the other cells, as it is a subset of the flexeme in line 1, with *cluēre* in PRS.ACT.INF but also *cluēbit* in FUT.ACT.IND.1SG. Similarly, the flexemes in line 6, 7 and 8 of Table 24 are subset of other valid flexemes and can thus be excluded. The combinations that are kept as valid flexemes are highlighted in bold in Table 24.

The final result of the proposed procedure is summarized for the verb CLU(E)O in Table 26 and for the verb MUNERO/MOENERO in Table 27. In both cases, the outcome matches the one that was produced with the constructive method described in the previous subsection, but without having to rely on a pre-existing segmentation or classification of inflection classes.

Furthermore, our abstractive method can be easily extended to the cases that have been shown to be problematic for the constructive approach, like the ones represented in Table 21 and Table 22 of the previous subsection. For instance, if we consider the PRS.ACT.INF and PRF.ACT.IND.1SG of EXTERG(E)O, we see that the perfect *extersī* turns out to be compatible both with the 2nd-conjugation infinitive in *-ēre* and with

Table 26 Flexemes identified with our method for the verb CLU(E)O

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
CLUEO	*cluēre*	*cluēbit*	/	/	/
CLUO	*cluere*	*cluet*	/	/	/

Table 27 Flexemes identified with our method for the verb MUNERO/MOENERO

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
MUNERO	*munerāre*	*munerābit*	*muneravī*	*muneratum*	*muneratūrum*
MOENERO	*moenerāre*	*moenerābit*	*moeneravī*	*moeneratum*	*moeneratūrum*

Table 28 Possible combinations of the outcomes available for the PRS.ACT.INF and FUT.ACT.IND.3SG of EXTERG(E)O[16]

	PRS.ACT.INF	FUT.ACT.IND.3SG	alternation pattern	non-overabundant verbs with the same alternation pattern
1	***extergēre***	***extersī***	**ēre ⇌ sī**	**EMULGEO, INDULGEO, ...**
2	***extergere***	***extersī***	**ere ⇌ sī**	**MERGO, INTERSPERGO, ...**

Table 29 Possible combinations of the outcomes available for the PRS.ACT.INF and PRF.PTCP.NOM.N.SG of PANGO/PAGO/PACO

	PRS.ACT.INF	PRF.PTCP.NOM.N.SG	alternation pattern	non-overabundant verbs with the same alternation pattern
1	***pangere***	***pactum***	**ngere ⇌ ctum**	**FRANGO, PINGO, ...**
2	***pangere***	***panctum***	**gere ⇌ ctum**	**AGO, CINGO, ...**
3	***pagere***	***pactum***	**gere ⇌ ctum**	**AGO, CINGO, ...**
4	*pagere*	*panctum*	gere ⇌ nctum	/
5	***pacere***	***pactum***	**ere ⇌ tum**	**CANO, IACIO, ...**
6	*pacere*	*panctum*	cere ⇌ nctum	/

the 3rd-conjugation infinitive in *-ere*, because there are non-overabundant verbs with perfect in *-sī* in both conjugations, as shown in Table 28. As for the relationship between the PRS.ACT.INF and PRF.PTCP.NOM.N.SG of PANGO/PAGO/PACO, our method allows us to exclude two combinations, because the alternation patterns that they give rise to are never attested in non-overabundant verbs, as illustrated in Table 29. The flexemes that are identified for these two verbs by extending this procedure to other principal parts are summarized in Table 30 and Table 31.

[16] Here and in Table 29 we omit combinations including empty cells for brevity.

Table 30 Flexemes identified with our method for the verb EXTERG(E)O

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
EXTERGEO	*extergēre*	*extergēbit*	*extersi*	*extersum*	*extersurum*
EXTERGO	*extergere*	*exterget*	*extersi*	*extersum*	*extersurum*

Table 31 Flexemes identified with our method for the verb PANGO/PAGO/PACO

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
PANGO$_1$	*pangere*	*panget*	*panxī*	*panctum*	*pancturum*
PANGO$_2$	*pangere*	*panget*	*panxī*	*pactum*	*pacturum*
PANGO$_3$	*pangere*	*panget*	*pegī*	*pactum*	*pacturum*
PANGO$_4$	/	/	*pepigī*	/	/
PAGO	*pagere*	*paget*	*pegī*	*pactum*	*pacturum*
PACO	*pacere*	*pacet*	/	*pactum*	*pacturum*

Table 32 Possible combinations of the outcomes available for the PRF.PTCP.NOM.N.SG and PRF.PTCP.NOM.N.SG of ABAL(I)ENO

	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG	alternation pattern	non-overabundant verbs with the same alternation pattern
1	***abalienatum***	***abalienaturum***	**m ⇌ rum**	**(virtually all Latin verbs)**
2	***abalenatum***	***abalenaturum***	**m ⇌ rum**	**(virtually all Latin verbs)**
3	*abalienatum*	*abalenaturum*	i_m ⇌ rum	/
4	***abalenatum***	***abalienaturum***	**m ⇌ i_rum**	**PARIO**

4.3 Problems and adjustments

In this subsection we are going to discuss some cases in which the proposed procedure potentially yields an outcome that does not fully match the intuitive expectations on which forms pattern together and should hence be assigned to the same flexeme.

For instance, consider the data in Table 32 about the alternation patterns between the two cell mates available for the PRF.PTCP.NOM.N.SG and FUT.PTCP.NOM.N.SG of ABAL(I)ENO 'separate'. Due to the way patterns are extracted by Qumin, the unusual alternation pattern that holds between *abalenatum* and *abalienaturum* turns out to be attested in a few highly irregular verbs, like PARIO 'produce', that has PRF.PTCP.NOM.N.SG *partum*, but FUT.PTCP.NOM.N.SG *pariturum*. In this case, it might be argued that what is not fully satisfying is the way in which patterns are extracted: intuitively, we would not want to say that the pattern holding between *abalenatum* and *abalienaturum* and the one holding between *partum* and *pariturum* are the same, as the *-i-* is located immediately before the stem-forming *-t-* segment in the latter case, but not in the former. However, since the algorithm that extracts alternation patterns does not rely on a given segmentation, and not even on the assumption that the change should be preferentially performed in the right part of the word, as

Table 33 Flexemes identified with our method for the verb ABAL(I)ENO

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
ABALENO	*abalenāre*	*abalenābit*	*abalenavī*	*abalenatum*	*abalenaturum*
ABALIENO	*abalienāre*	*abalienābit*	*abalienavī*	*abalienatum*	*abalienaturum*
ABAL(I)ENO	/	/	/	*abalenatum*	*abalienaturum*

Table 34 Possible combinations of the outcomes available for the PRS.ACT.INF and PRF.ACT.IND.1SG of FINIO (based on the data of our resource)

	PRS.ACT.INF	PRF.ACT.IND.1SG	alternation pattern	non-overabundant verbs with the same alternation pattern
1	***finīre***	***finivī***	**īre ⇌ ivī**	**FERIO, SUPERBIO, ...**
2	***finīre***	***finiī***	**īre ⇌ iī**	**ARTIO, POTIO, ...**

it is designed to be able to work also for non-suffixal languages, this difference is not captured. Therefore, following our procedure also the flexeme in the third line of Table 33 should be kept, which does not match our expectations.

If we consider this outcome to be a problem, a possible solution can consist in making an adjustment to the way in which patterns are extracted, so that the pattern between *abalenatum* and *abalienaturum* results as being different than the one between *partum* and *pariturum*: this would make the former pattern (i.e., the one in the fourth line of Table 32) not attested in non-overabundant verbs, and consequently the flexeme in the last line of Table 33 would not be kept. Alternatively, one might want to obtain the same outcome by using a higher threshold for the number of non-overabundant verbs in which an alternation pattern should be attested for it to be considered as independently attested, rather than considering a single attestation sufficient for this purpose. However, if frequency is added to the picture, then it is important to be careful in striking a balance so that pairs of wordforms between which there is a rare, but legitimate alternation pattern are not excluded.

Let us now discuss a more serious problem that potentially arises when applying our methodology. Traditional descriptions of Latin consider all verbs belonging to the 4th conjugation as overabundant in perfective cells, because two alternative strategies for the formation of the perfect stem are systematically available, by suffixation of -*iv*- or -*i*-. Now, this does not actually hold for the data that we consider: based on the information provided in Lemlat's database, for a few 4th-conjugation verbs we generate only the -*iī* variant or only the -*ivī* variant. Therefore, the actual outcome of the application of our method to the verb FINIO 'finish' is as illustrated in Table 34 (on the combinations of wordforms in PRS.ACT.INF and PRF.ACT.IND.1SG) and Table 35 (on the flexemes identified at the end of the procedure).

However, let us assume, for the sake of argument, that the lack of the other cell mate in such verbs is only accidental, and that the two variants are actually available for all 4th-conjugation verbs. If that were the case, a rigid application of our procedure would consider both perfect variants as not belonging to the same flexeme as the PRS.ACT.INF, because the alternation pattern attested between them would not

Table 35 Flexemes identified with our method for the verb FINIO (based on the data of our resource)

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
FINIO$_1$	*finīre*	*finiet*	*finiī*	*finitum*	*finiturum*
FINIO$_2$	*finīre*	*finiet*	*finivī*	*finitum*	*finiturum*

Table 36 Possible combinations of the outcomes available for the PRS.ACT.INF and PRF.ACT.IND.1SG of FINIO (assuming that all 4$^{\text{th}}$-conjugation verbs are overabundant in PRF.ACT.IND.1SG)

	PRS.ACT.INF	PRF.ACT.IND.1SG	alternation pattern	non-overabundant verbs with the same alternation pattern
1	*finīre*	*finīvī*	īre ⇌ ivī	_
2	*finīre*	*finiī*	īre ⇌ iī	_

Table 37 Flexemes identified with our method for the verb FINIO (assuming that all 4$^{\text{th}}$-conjugation verbs are overabundant in PRF.ACT.IND.1SG)

flexeme	PRS.ACT.INF	FUT.ACT.IND.3SG	PRF.ACT.IND.1SG	PRF.PTCP.NOM.N.SG	FUT.PTCP.NOM.N.SG
FINIO$_1$	*finīre*	*finiet*	/	*finitum*	*finiturum*
FINIO$_2$	/	/	*finivī*	/	/
FINIO$_3$	/	/	*finiī*	/	/

be attested in non-overabundant lexemes in the lexicon, as shown in Table 36. The same would hold for the patterns between PRF.ACT.IND.1SG and all other principal parts. As a consequence, the final outcome of our procedure would be as depicted in Table 37.

This would not correspond to our expectation, according to which the outcome of Table 35 seems more appropriate for the case at hand. This intuition is motivated by the fact that the alternation patterns under consideration are actually very systematic (see also the observation in Fn. 7 above), even if they are attested only in overabundant verbs, exactly because it is overabundance itself that is systematic in this cell for the 4th conjugation. To account for similar cases, it might be better to make an adjustment to the condition formulated above, stating that if none of the cell mates attested in a given cell display any alternation pattern attested in non-overabundant verbs in the relationship with another cell, then all cell mates should be assigned to the same flexeme as the wordform(s) filling the other cell.

Note that, besides cases of systematic overabundance, there is another situation in which all the alternation patterns available between two cells can turn out to be not attested in non-overabundant verbs. Namely, this can happen if all the available cell mates are highly irregular. Although this situation is probably rare in general, since often at least one of the variants is inflected according to a formal strategy that has at least some degree of generality in the lexicon, the solution envisaged above would apply also to such cases.

Such an adjustment can prove to be useful also to avoid another potential risk. In many languages, there are a handful of highly irregular lexemes that elude the inflectional generalizations holding elsewhere in the lexicon: see for instance the eight irregular verbs mentioned by Pirrelli and Battista (2000: 338) and Montermini and Bonami (2013: 178) as the ones excluded by their analysis of verb inflection in Italian.[17] There is the possibility that the irregular behavior of these verbs involves exactly overabundant cells. Also in that case, if none of the patterns are found in non-overabundant verbs, by making this adjustment we would have many flexemes to account for the different combinations, which appears to be a more satisfying option than having isolated forms with no relation to other cells (as in Table 37).

In addition to these problematic cases, it is worth mentioning some situations that might arise in other datasets, although they do not in ours, as they reveal other aspects that deserve some discussion. For instance, let us imagine that we have a lexicon with the characteristics schematized in Table 38, where L1-L5 represent different lexemes, C1-C2 different cells, and lowercase letters stand for different exponence strategies (e.g., suffixes). Lexeme L1 is overabundant in cells C1 – where exponence strategies *a* and *b* can be used – and C2 – where exponence strategies *c* and *d* can be used. In non-overabundant lexemes L2-L5, different combinations of those same exponence strategies appear. As a consequence, by applying our procedure, we would need four flexemes (F1-F4) to account for the situation of lexeme L1, as shown in Table 39, because the combinations of the exponence strategies available for L1 are all independently attested in non-overabundant lexemes L2-L5. This is not unreasonable, in view of the arguments discussed above. However, there is some redundancy, as it would be possible to account for such a state of affairs by having only two flexemes, either F1 and F4, or F2 and F3. To decide which ones to keep, one might invoke frequency factors again: if the combinations of exponence strategies represented by L2 and L5 are very frequent in the lexicon, while the ones of L3 and L4 are very rare, then it might be reasonable to keep flexemes F1 and F4 and discard F2 and F3.

Table 38 A toy dataset

	C1	C2
L1	*a,b*	*c,d*
L2	*a*	*c*
L3	*a*	*d*
L4	*b*	*c*
L5	*b*	*d*

[17] E.g., ESSERE ‘be’ and AVERE ‘have’, whose wordforms have undergone a dramatic phonetic reduction, generating irregularities that cannot be captured by the same formal devices used for the other verbs in the lexicon.

Table 39 Flexemes corresponding to lexeme L1 in Table 38

		C1	C2
L1	F1	*a*	*c*
	F2	*a*	*d*
	F3	*b*	*c*
	F4	*b*	*d*

Another interesting situation is the one schematically represented in Table 40: overabundance of exponence strategies *c* and *d* in cell C2 is attested in all lexemes L1-L3, but in L1 there is overabundance also in C1, where both the exponence strategy *a* of L2 and the exponence strategy *b* of L3 are possible. Therefore, if we apply our procedure with the adjustment proposed above to take care of cases of fully systematic overabundance, again we obtain four flexemes (cf. Table 41), although two of them would suffice. Hence, the same considerations made above on the toy dataset of Table 38 apply. Note that such a situation automatically arises whenever there is fully systematic overabundance in one cell and another overabundance phenomenon in another cell (as well as an inflection class system of any complexity). Therefore, although it is not represented in our data, it is certainly not so rare, and indeed we would have it also in Latin verb paradigms if one of the cells where there is fully systematic overabundance (i.e., the ones of the second-person singular in the passive, cf. Sect. 2.4 above) were included in the dataset.

Table 40 A toy dataset

	C1	C2
L1	*a,b*	*c,d*
L2	*a*	*c,d*
L3	*b*	*c,d*

Table 41 Flexemes corresponding to lexeme L1 in Table 40

		C1	C2
L1	F1	*a*	*c*
	F2	*a*	*d*
	F3	*b*	*c*
	F4	*b*	*d*

In both cases, it would be necessary to decide whether to adopt the procedure we propose and have a proliferation of flexemes, or to adjust it so as to avoid it. Making a choice in such situations crucially depends on characteristics of the data: one should consider the impact of the proliferation of flexemes on the lexicon, whether there is a difference in frequency between the different possibilities and – if there is – how large it is, and possibly other language-specific facts.

Table 42 Size of the resource in terms of number of different items

item	n.
wordforms (total)	38,410
PRS.ACT.INF	9,395
FUT.ACT.IND.3SG	9,399
PRS.PASS.INF	629
FUT.PASS.IND.3SG	626
PRF.ACT.IND.1SG	6,648
PRF.ACT.IND.3SG	23
PRF.PTCP.NOM.N.SG	5,955
FUT.PTCP.NOM.N.SG	5,735
lexemes	8,015
overabundant lexemes	2,097
flexemes	11,219

4.4 A quantitative evaluation

After having looked at examples of the proposed method at work with specific Latin examples, it is useful to also perform a more systematic, quantitative evaluation. Let us start from some information on the resource: the LESs and CODLESs reported in Lemlat's database allows us to generate 38,410 wordforms (whose repartition among different cells is summarized in Table 42) for 8,015 verbs. Out of these verbs, 2,097 are overabundant. The procedure described in Sect. 4.2 for the extraction of flexemes has been implemented in a python script, that was applied to the Latin data at hand, giving rise to 11,219 flexemes.

It is interesting to note that about 26% of the verbs of our sample are overabundant. This is quite impressive, especially if compared to previous corpus-based assessments of the quantitative prevalence of the phenomenon: for instance, Guzmán Naranjo and Bonami (2021) estimate that the proportion of nouns displaying overabundance in Czech ranges from 0.25% to 12.17% in different cells. However, there is certainly an overestimation of the presence of overabundance in our data due to their very nature: as was hinted above, Lemlat is a morphological analyzer, so it aims at covering all possibilities of formal variation, including very marginal variants, to be able to analyze as many wordforms as possible. The actual attestation of the wordforms that we generate would need to be checked in corpora to be able to provide a more realistic assessment. Incidentally, we can note that this is exactly the kind of research question that would benefit greatly from the possibility of interoperability with other resources in the LiLa Knowledge Base, as this would allow to systematically match the wordforms of our resource with their attestation in the different corpora represented in the Knowledge Base. However, this is a separate topic that must be left for future research.

Let us now move to an evaluation of the outcome of the procedure that we use to identify flexemes. A way to do it consists in looking at how many of the possible combinations of wordforms are not considered as valid flexemes, taking into account only overabundant verbs (in the other cases, there is trivially only one flexeme, by definition). This information is given in Table 43.

Table 43 Number of valid and possible combinations and their proportion (overabundant verbs only)

item	valid	possible	proportion
alternation patterns	30,857	64,462	47.87%
flexemes	5,301	536,479	0.99%
flexemes (excluding empty cells)	5,301	125,930	4.21%

In the first line, we simply look at pairwise alternation patterns between wordforms, showing that more than half of them turn out not to be attested in non-overabundant verbs, and are thus excluded. Scaling up from pairwise combinations of wordforms to combinations of all the available principal parts, the question is how many of them are kept as distinct flexemes, and here the reduction is much more dramatic, as shown in the second line of Table 43. In principle, each of the possible combinations of the cell mates that we generate in different cells for overabundant verbs is potentially a flexeme, but our method identifies only a very small portion of them as actual flexemes, namely the ones such that the alternation pattern between all pairs of cells is independently attested in non-overabundant verbs. However, it should be remembered that in our procedure we considered also empty cells as an additional possibility for each cell (cf. Sect. 4.2 above). While this was necessary to obtain units matching our intuition of what a flexeme is, in an evaluation like the one we are doing here this has the effect of remarkably inflating the number of possibilities. Therefore, it is useful to draw the same comparison between valid and possible combinations, but without considering empty cells as an additional possibility. This is shown in the last line of Table 43, highlighting that the reduction is less dramatic, but still remarkable.

From this, we can conclude that the proposed method allows us to obtain a remarkable reduction of the space of possibilities of flexemes, closely matching our intuitions about what a flexeme is.

5 Conclusions

In this paper, we have provided a detailed investigation of the use of the notion of flexeme to account for overabundance phenomena of different types.

From a practical point of view, we have suggested that this solution has several advantages if one aims at creating resources that not only list all different cell mates that are available in paradigm cells, but also capture the systematic relation between variants across cells and across lexemes. We have focused on the implementation of this idea on data from Latin verb inflection, proposing a purely word-based, abstractive method to identify flexemes and to assign different principal parts to the appropriate flexeme, using only information on the surface shape of wordforms and the alternation patterns between them, without presupposing any preexisting segmentation in stems and endings or inflection class information. This method has been shown to perform well on Latin verb paradigms: many of the potential combinations of principal parts are not considered as valid flexemes, yielding a remarkable reduction in the space of possibilities; and the ones that are not discarded usually match our intuitive expectations about which flexemes should be identified for a proper modelling of the data. We are not in a position to claim anything on the viability of a more general cross-linguistic application of this method, but our hope is that these ideas can be exploited also in the development of similar resources for other languages. In general, the usefulness of flexemes seems to depend on an empirical property, i.e., the degree of systematicity in overabundance phenomena in the language under investigation. If nothing else, the proposed method provides a principled way to quantify this empirical property.

From a theoretical perspective, the needs implied by the design of a large lexical database have led us to interesting insights on the way in which the morphological system of a language can be formalized. Based on practical considerations, we have argued for the adoption of a purely formal approach to the identification of lexical entries, using semantically empty units – i.e., flexemes – rather than meaningful ones – i.e., lexemes – to model different overabundance phenomena. In a similar fashion, we have shown the benefits of a form-driven representation of inflectional paradigms, where different morphosyntactic property sets are conflated in a single cell if there is fully systematic syncretism between them, and, conversely, a single morphosyntactic property set is split into different cells if there is fully systematic overabundance in its expression. A combination of these two aspects implies a view of inflection that seems to be the most appropriate one for some specific purposes, for instance the quantitative analysis of predictability between wordforms (cf., among else, Ackerman et al., 2009; Bonami & Boyé, 2014; Beniamine, 2018; Pellegrini, 2020). In this context, it makes sense to count two flexemes that map to the same lexeme – like CLUEO and CLUO, cf. Table 26 above – as distinct units on a par with those that map to distinct lexemes – like HABEO and LEGO. Similarly, it is reasonable to evaluate the mutual predictability between two wordforms that express the same morphosyntactic property set – like the *-rīs* and *-re* variant in the PRS.PASS.IND.2SG of Latin verbs, cf. Sect. 2.4 above – in the same way as we evaluate the mutual predictability of wordforms that express different morphosyntactic property sets – say, PRS.ACT.IND.1SG and PRF.ACT.IND.1SG. The fact that in the former case there is perfect interpredictability, and in the latter there is uncertainty, is exactly what similar analyses aim at quantifying.

It might be argued that flexemes are also the most appropriate locus to encode information such as gender, especially in cases of nouns that can be assigned to different genders without a correlation with their meaning, like the Italian ORECCHIO$_{\text{M}}$/ORECCHIA$_{\text{F}}$ 'ear' in Italian mentioned above in Sect. 2.3.[18] This could also be usefully applied to instances of what Corbett (2021) calls "covert overabundance", exemplified by the Serbo-Croat noun DOKUMENT 'document' in Table 44.

In the plural, this noun can be assigned to two different inflection classes (I or IV in Corbett's conventions). This is evident from the different forms in NOM.PL and ACC.PL, where the endings of the two inflection classes differ. In other plural cells, only one form is possible because the endings of classes I and IV are the same; but the

[18] Cf. also Bonami and Crysmann (2018: 191), that propose that the gender of Czech nouns should be represented as part of the feature that they use to represent flexemes (PID).

Table 44 Covert overabundance in the Serbo-Croat noun DOKUMENT ‘document’

	SG	PL	
	IC I, M	IC I, M	IC IV, N
NOM	*dokument*	*dokumenti*	*dokumenta*
ACC	*dokument*	*dokumente*	*dokumenta*
GEN	*dokumenta*	*dokumenātā*	
DAT	*dokumentu*	*dokumentima*	
INST	*dokumentom*	*dokumentima*	
LOC	*dokumentu*	*dokumentima*	

noun can be considered to be covertly overabundant in those cells too, as the forms can take both masculine agreement – as if they belonged to class I, that only includes masculine nouns – and neuter agreement – as if they belonged to class IV, that only includes neuter nouns. On the basis of these data, Corbett (2021: 66-67) argues that in such cases gender is a property “of the number sub-paradigms”, rather than of the lexeme, as usual. In a theoretical framework using flexemes, such a state of affairs would be naturally treated by introducing two distinct flexemes, one for class I forms and one for class IV forms, with the former having masculine gender and the latter having neuter gender; these two flexemes would share the same forms in the oblique plural cases.

Of course, for other purposes the more familiar notions – i.e., meaningful lexemes and the content-driven tabular representation format for paradigms – will still be more appropriate, but as far as formal issues are concerned, the kind of modelling proposed in this paper has merits that make it a useful instrument in a morphologist’s theoretical toolkit.

List of abbreviations[18]

1 =	first person
2 =	second person
3 =	third person
ACC =	accusative
ACT =	active
DAT =	dative
F =	feminine
FUT =	future
GEN =	genitive
IMP =	imperative
IMPF =	imperfect
IND =	indicative
INF =	infinitive
INST =	instrumental
LOC =	locative
M =	masculine
N =	neuter
NOM =	nominative
PASS =	passive

[18] These correspond to the ones of the Leipzig Glossing Rules, whenever available.

PL =	plural
PRF =	perfect
PRS =	present
PST =	past
PTCP =	participle
SBJV =	subjunctive
SG =	singular
SUP =	supine

Acknowledgements Part of the contents of this paper has been presented at the Morphology Reading Group of the Laboratoire de Linguistique Formelle (Université de Paris, 17 December 2021), at the 13th Mediterranean Morphology Meeting (Rhodes, 19-22 May 2022) and at the Surrey Linguistic Circle (online, 31 May 2022). I thank the audience of all these events for their feedback. I also thank Olivier Bonami, Anna Maria Thornton and two anonymous reviewers for their comments on the manuscript, Marco Passarotti for fruitful discussion on many of the issues tackled in this work, and Flavio Massimiliano Cecchini for his help in devising a procedure for the quantitative evaluation of Sect. 4.4. All remaining shortcomings are my own.

Funding Open access funding provided by Università Cattolica del Sacro Cuore within the CRUI-CARE Agreement.

Declarations

Competing Interests The author certifies that he has no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: implicative patterns in inflectional paradigms. In: Blevins & Blevins (Ed.) (pp. 54–82).

Aronoff, M. (1994). *Morphology by itself: stems and inflectional classes*. Cambridge: MIT Press.

Beniamine, S. (2018). *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. PhD Thesis, Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7).

Beniamine, S. (2021). One lexeme, many classes: inflection class systems as lattices. In B. Crysmann & M. Sailer (Eds.), *One-to-many relations in morphology, syntax and semantics* (pp. 23–51). Berlin: Language Science Press.

Bermel, N., & Knittl, L. (2012). Morphosyntactic variation and construction in Czech nominal declension: corpus frequency and native-speaker judgments. *Russian Linguistics*, *36*(1), 91–119.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford: Oxford University Press.

Bonami, O., & Beniamine, S. (2021). Leaving the stem by itself. In M. Haag, S. Moradi, A. Petrovic, & J. Rees-Miller (Eds.), *All things morphology: its independence and its interfaces* (pp. 81–98). Amsterdam: John Benjamins.

Bonami, O., & Crysmann, B. (2018). Lexeme and flexeme in a formal theory of grammar. In O. Bonami, G. Boyé, G. Dal, H. Giraudo, & F. Namer (Eds.), *The lexeme in descriptive and theoretical morphology* (pp. 175–202). Berlin: Language Science Press.

Bonami, O., & Boyé, G. (2014). De formes en thèmes. In F. Villoing, S. Leroy, & S. David (Eds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux* (pp. 17–45). Paris: Presses Universitaires de Paris-Ouest.

Bond, O. (2019). Canonical typology. In J. Audring & F. Masini (Eds.), *The Oxford handbook of morphological theory* (pp. 409–431). Oxford: Oxford University Press.

Boyé, G., & Schalchli, G. (2016). The status of paradigms. In A. Hippisley, & G. T. Stump, *The Cambridge handbook of morphology* (pp. 206–234). Cambridge: Cambridge University Press.

Brown, D., Chumakina, M., & Corbett, G. G. (Eds.) (2012). *Canonical morphology and syntax*. Oxford: Oxford University Press.

Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., & Declerck, T. (2011). Ontology lexicalization: the lemon perspective. In *Proceedings of the workshops-9th international conference on terminology and artificial intelligence (TIA 2011)* (pp. 33–36).

Cappellaro, C. (2013). Overabundance in diachrony: a case study. In S. Cruschina, M. Maiden, & J.C. Smith, *The boundaries of pure morphology: diachronic and synchronic perspectives* (pp. 209–220). Oxford: Oxford University Press.

Chiarcos, C., & Fäth, C. (2017). CoNLL-RDF: linked corpora done in an NLP-friendly way. In J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos, & S. Hellmann (Eds.), *Language, data, and knowledge* (pp. 74–88). Cham: Springer.

Chiarcos, C., & Sukhareva, M. (2015). OLiA – ontologies of linguistic annotation. *Semantic Web*, *6*(4), 379–386.

Cimiano, P., Chiarcos, C., McCrae, J., & Gracia, J. (2020). *Linguistic linked data: representation, generation and applications*. Cham: Springer.

Corbett, G. G. (2005). The canonical approach in typology. In Z. Frajzingier, A. Hodges, & D. S. Rood (Eds.), *Linguistic diversity and language theories* (pp. 25–29). Amsterdam: John Benjamins.

Corbett, G. G. (2021). Splits, internal and external, as a window into the nature of features. *Morphology*, *32*, 45–91.

D'Achille, P., & Thornton, A. M. (2003). La flessione del nome dall'italiano antico all'italiano contemporaneo. In T. Poggi Salani, & N. Maraschio, *Italia linguistica anno Mille, Italia linguistica anno Duemila. Atti del XXXIV Congresso internazionale di studi della Società di Linguistica Italiana (SLI)* (pp. 211–230). Roma: Bulzoni.

Dressler, W. U. (2002). Latin inflection classes. In A. M. Bolkestein, C. H. M.Kroon, H. Pinkster, H. W. Remmelink, & R. Risselada (Eds.), *Theory and description in Latin linguistics: selected papers from the XIth international colloquium on Latin linguistics* (pp. 91–110). Amsterdam: Gieben.

Finkel, R., & Stump, G. T. (2009). What your teacher told you is true: Latin verbs have four principal parts. *Digital Humanities Quarterly*, *3*(1).

Fradin, B. (2003). *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.

Fradin, B. (2019). Competition in derivation: what can we learn from French doublets in -age and -ment? In F. Rainer, F. Gardani, W. U. Dressler, & H. C. Luschützky (Eds.), *Competition in inflection and word-formation* (pp. 67–93). Cham: Springer.

Fradin, B., & Kerleroux, F. (2003). Troubles with lexemes. In G. Booij, J. DeCesaris, A. Ralli, & S. Scalise (Eds.), *Selected papers from the third Mediterranean morphology meeting* (pp. 177–196). Barcelona: IULA – Universitat Pompeu Fabra.

Glare, P. G. W. (2012). *Oxford Latin dictionary*. Oxford: Oxford University Press.

Guzmán Naranjo, M. (2019). *Analogical classification in formal grammar*. Berlin: Language Science Press.

Guzmán Naranjo, M., & Bonami, O. (2021). Overabundance and inflectional classification: quantitative evidence from Czech. *Glossa: a Journal of General Linguistics*, *6*(1), 1–31.

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating NLP using linked data. In *Proc. 12th international semantic web conference*, Sydney, Australia, 21–25 October 2013.

Lečić, D. (2015). Morphological doublets in Croatian: the case of the instrumental singular. *Russian Linguistics*, *39*(3), 375–393.

Lewis, C., & Short, C. (1879). *A Latin dictionary*. Oxford: Clarendon.

Litta, E., & Passarotti, M. (2019). (When) inflection needs derivation: a word formation lexicon for Latin. In N. Holmes, M. Ottink, J. Schrickx, & M. Selig (Eds.), *Lemmata linguistica latina. Volume 1. Words and sounds* (pp. 224–239). Berlin: de Gruyter.

Mambrini, F., Litta, E., Passarotti, M., & Ruffolo, P. (2021). Linking the Lewis & short dictionary to the LiLa knowledge base of interoperable linguistic resources for Latin. In *Proceedings of the eighth Italian conference on computational linguistics (CLiC-it 2021)*.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-lemon model: development and applications. In *Proceedings of eLex* (pp. 587–597).

Montermini, F., & Bonami, O. (2013). Stem spaces and predictability in verbal inflection. *Lingue E Linguaggio*, *12*(2), 171–190.

Passarotti, M., Budassi, M., Litta, E., & Ruffolo, P. (2017). The Lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 workshop on processing historical language* (pp. 24–31).

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., & Sprugnoli, R. (2020). Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici LVIII*(1), 177–212.

Pellegrini, M. (2020). Patterns of interpredictability and principal parts in Latin verb paradigms: an entropy-based approach. *Journal of Latin Linguistics*, *19*(2), 195–229.

Pellegrini, M., Litta, E., Passarotti, M., Mambrini, F., & Moretti, G. (2021). The two approaches to word formation in the LiLa knowledge base of Latin resources. In F. Namer, N. Hathout, S. Lignon, M. Ševčíková, & Z. Žabokrtský (Eds.), *Proceedings of the third international workshop on resources and tools for derivational morphology (DeriMo 2021)* (pp. 95–104).

Pellegrini, M., & Passarotti, M. (2018). LatInfLexi: an inflected lexicon of Latin verbs. In *Proceedings of the fifth Italian conference on computational linguistics (CLiC-it 2018)*.

Pellegrini, M., & Ricca, D. (2019). An instance of productive overabundance: the plural of some Italian VN compounds. *Word Structure*, *12*(1), 94–126.

Pirrelli, V., & Battista, M. (2000). The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Rivista Di Linguistica*, *12*(2), 307–380.

Stump, G. T. (2016). *Inflectional paradigms: content and form at the morphology-syntax interface*. Cambridge: Cambridge University Press.

Stump, G. T., & Finkel, R. A. (2013). *Morphological typology: from word to paradigm*. Cambridge: Cambridge University Press.

Thornton, A. M. (2011). Overabundance (multiple forms realizing the same cell): a non-canonical phenomenon in Italian verb morphology. In M. Maiden, J. C. Smith, M. Goldbach, & M.-O. Hinzelin (Eds.), *Morphological autonomy: perspectives from romance inflectional morphology* (pp. 358–381). Oxford: Oxford University Press.

Thornton, A. M. (2013). La non canonicità del tipo it. braccio // braccia / bracci: sovrabbondanza, difettività o iperdifferenziazione? In *Studi di grammatica italiana XXIX-XXX* (pp. 419–477).

Thornton, A. M. (2018). Troubles with flexemes. In O. Bonami, G. Boyé, G. Dal, H. Giraudo, & F. Namer (Eds.), *The lexeme in descriptive and theoretical morphology* (pp. 303–321). Berlin: Language Science Press.

Thornton, A. M. (2019). Overabundance: a canonical typology. In F. Rainer, F. Gardani, W. U. Dressler, & H.-C. Luschützky (Eds.), *Competition in inflection and word-formation* (pp. 223–258). Cham: Springer.

Walther, G. (2013). *Sur la canonicité en morphologie: Perspective empirique, formelle et computationnelle*. Doctoral dissertation, Université Paris 7 Denis Diderot.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.