Non-Canonical Inflection: Data, Formalisation and Complexity Measures

Benoît Sagot, Géraldine Walther

Non-Canonical Inflection: Data, Formalisation and Complexity Measures

Benoît Sagot* and Géraldine Walther**

* ALPAGE, INRIA Paris–Rocquencourt & Université Paris 7
** Univ. Paris Diderot, Sorbonne Paris Cité, LLF, UMR 7110, 75013, Paris, France
benoit.sagot@inria.fr,geraldine.walther@linguist.jussieu.fr

Abstract. Non-canonical inflection (suppletion, deponency, heteroclisis...) is extensively studied in theoretical approaches to morphology. However, these studies often lack practical implementations associated with large-scale lexica. Yet these are precisely the requirements for objective comparative studies on the complexity of morphological descriptions. We show how a model of inflectional morphology which can represent many non-canonical phenomena [67], as well as a formalisation and an implementation thereof can be used to evaluate the complexity of competing morphological descriptions. After illustrating the properties of the model with data about French, Latin, Italian, Persian and Sorani Kurdish verbs and about noun classes from Croatian and Slovak we expose experiments conducted on the complexity of four competing descriptions of French verbal inflection. The complexity is evaluated using the information-theoretic concept of *description length*. We show that the new concepts introduced in the model by [67] enable reducing the complexity of morphological descriptions w.r.t. both traditional or more recent models.

Keywords: Inflectional Morphology, Description Complexity, MDL, Paradigm Shape, Canonicity, Inflection Zone, Stem Zone, Inflection Pattern, Stem Pattern.

1 Introduction

Automatically generating all forms of a language's inflectional paradigms is often considered a rather unchallenging task, since it has long been solved for most languages of interest to the area of natural language processing (NLP).

On the other hand, there is much ongoing work in theoretical morphology within lexicalist approaches, and especially within *Word and Paradigm* related frameworks [46, 72, 1, 60, 26], on describing, modeling and explaining inflection, and in particular non-canonical inflection phenomena. For example, the Surrey Morphology Group has been working on projects on *Syncretism* (1999-2002), *Suppletion* (2000-2003), *Deponency* (2004-2006) and *Defectiveness* (2006-2009). In 2003 G. G. Corbett publishes his first article on *Canonical Typology* [23], thereby laying the foundations for a theoretical approach aiming at capturing the discrepancy between regularity and irregularity in inflectional paradigms.

However, studies in theoretical morphology are sometimes limited by the lack of complete formalisations and large-scale implementations of the concepts they manipulate, both in terms of morphological and lexical coverage. Still, such resources are required for achieving qualitative assessments of the validity of a given approach or to compare the relevance of several morphological models describing a given language or a specific part of a language's morphology, including the information encoded in the lexicon. This direction of research points towards recent work aiming at measuring linguistic and more specifically morphological *complexity* [7]. Indeed, this provides valuable insights into typological phenomena and properties of linguistic structures, and allows to compare various linguistic descriptions with objective metrics (see Section 5).

In this paper, we follow a *Word and Paradigm* based view of morphology. We introduce metrics allowing for measuring the complexity of a morphological description. The underlying idea is that inflectional complexity lies in the amount and distribution of inflectional irregularities. Irregularities can be represented as specific rules within the morphological grammar or as additional information within the lexicon. Hence our complexity metrics apply to both the morphological grammar and the information stored in the lexicon, thus allowing for comparing different competing descriptions in terms of descriptive complexity.

After a brief summary of the related work in both computational and theoretical inflectional morphology (Section 2), we first recall the definition of a large variety of non-canonical inflectional phenomena likely to cause increased descriptive complexity. These definitions are illustrated with data from French, Latin, Italian, Sorani Kurdish, Persian, Croatian and Slovak (Section 3). In Section 4, we then present our formal model of inflectional morphology.[1] We show how it covers all those non-canonical phenomena and allows for a formal representation of inflectional irregularity. Finally, in Section 5, we implement our model, putting a particular emphasis on French verbal inflection, which exhibits several of these phenomena and received much renewed attention in the last few years. We show how the implementation of this formal model within the Alexina lexical framework [56] makes it possible to define an information-theoretic notion of complexity for morphological descriptions that includes both the model and the corresponding lexicon, based on *description length*. We assess the complexity of four different accounts of French verbal inflection. As a side-result of this experiment, we also show that our formal model is not only able to encode previously proposed morphological descriptions [15, 56] but also provides a way to write a description of French verbal inflection that has a lower complexity.

2 Related Work

2.1 Related Work in Computational Morphology

Within contemporary computational morphology, inflection is treated in rule-based, supervised and unsupervised methods, sometimes combined [62].

[1] The reader will find a complete formal presentation in [67].

Among the first are : (i) stemming methods like the desuffixation algorithm by Porter [53]; (ii) bi-directional analysis and generation methods like Koskeniemmi's *Two-level morphology* [42] that uses transducers linking a deep (lexical) level to surface forms by applying systematic phonotactic transformations, as well as other finite state approaches [9]; (iii) morphosemantic approaches, mostly for specialised corpora [54, 45, 50, 21].

The second type of approach, namely supervised learning methods, rely on annotated learning corpora that define the expected output [16, 59].

Finally, there are the unsupervised approaches which can be used even for languages for which no preliminary description is available. These approaches rely on at least four different methods.

- Acquisition of morphological information can be achieved through direct comparison of graphemes distributed over a given corpus. This has be done through *edit distance measures* [8], maximum affix recognition [37, 31, 71], word insertion tests [39, 28, 10], and analogies [43, 35, 49];
- Another method relies on entropy models based on Harris' hypothesis [34]. They are mainly used for automatically detecting morph(eme)-boundaries through entropy measures [39, 10, 58];
- Probabilistic methods are of various types: bayesian inference models [27], bayesian hierarchy models [57] or probabilistic generative models [58].
- Segmentation methods relying on data compression [32, 27] using the *Minimum Description Length (MDL)* principle [55]. The underlying idea is that morphology always tends to use the most compact encoding by relying on inflectional regularity (see Section 5 for more details). Such regularity is stated to show in the stem vs. exponent [46] distribution, which should allow for a segmentation of corpora into the smallest possible morph(eme)-units;

All these unsupervised methods can also be combined with rule-based models, as in Tepper and Xia [62] who define contextual re-writing rules which they apply to the results of an unsupervised analysis in order to account for allomorphy in English and Turkish.

Although there appear to be many different methods that can be used in Computational Morphology, none of these explicitly tackle the question of complexity, regularity or canonicity *per se*.

2.2 Related Work in Theoretical Morphology

Morphological complexity however plays an important role in modern theoretical approaches to morphology. Within formal approaches to morphology, there are those who accept the existence of morphology and those who refuse it. The latter approaches are quite widespread and represented by Chomsky and Halle [22], Lieber [44] and *Distributed Morphology*, Halle and Marantz [33], while the former are illustrated by what is called the *Word and Paradigm* approach, e.g. Matthews, Aronoff and Stump [46, 2, 60]. Only in the *Word and Paradigm* approach, which are lexeme-based, does the question whether there is regularity

within paradigms really matter. In this work, we consequently adopt a lexeme-based approach to morphology that vastly relies on Matthews' view of stems and exponents [46].

The question of regularity in morphology is not always specifically addressed. Even in lexeme-based approaches, some works do not give the notion of *regularity* any theoretical status [60]. Yet, based on psycholinguistic evidence, such as presented by Pinker [51], there are modern approaches, such as illustrated in *Deriving Inflectional irregularity* by Bonami and Boyé [13], that treat *irregularity* as "a real grammatical phenomenon, that is manifest not only in psycholinguistic behaviour but also in language change and in synchronic grammar".

Work on (ir-)regularity in lexeme-based approaches has also been done in the subfield of *Canonical Typology*, such as presented by Corbett [23, 25]. Non-canonical/irregular phenomena such as *suppletion* [19, 13], *deponency* [6], *heteroclisis* [61], *defectiveness* [5] and more recently *overabundance* [64] have been studied within this approach, giving rise to quite a series of publications.[2]

However, these works have seldom explicitly targeted the development of descriptions that optimise their compacity. Indeed, in order to be able to evaluate a description's compacity, a large scale implementation is required. This implementation has to rely on large-scale lexical resources covering (almost) all the described language's relevant lexical items. It must also be able to implement the measured descriptions. A short state-of-the-art presentation of existing compacity measures is given in the introduction to Section 5.

3 Data on Non-Canonical Inflection

Couching our work in a *Word and Paradigm* approach to morphology, we define a morphological description as the combination of a set of inflectable lexical entries and corresponding realisation rules realising specific morphosyntactic features. The result of all applied realisation rules are the paradigms of a language's lexemes. In order to assess the complexity of specific morphological descriptions, we start with identifying those phenomena that tend to be paradigm-complexity-increasing. These phenomena are the irregular, non-default cases. In terms of *Canonical Typology*, they are the non-canonical inflectional phenomena.

3.1 Canonical Inflection

The concept of *canonical typology* can be traced back to Corbett [23] in an attempt to better understand what exactly differs from a hypothetic ideal *canonical* stage in the different occurrences of non-canonical phenomena. In this approach, *canonical inflection* must not to be mistaken for *prototypical inflection*. Canonical inflection is rare. It corresponds to an ideal stage, seldom met, but that

[2] Existing work is mostly done on phonological data. Our work focuses on written data for now. We plan on doing some future work on comparing phonological and graphemic morphological complexity.

constitutes a purely theoretical space from which deviant phenomena can be formally distinguished [24].

Canonical inflection is a notion that affects both the relation between the cells of a given lexeme's paradigm and the corresponding cells belonging to two different lexemes' paradigms. Canonical inflection is thus defined through the comparison of both the cells of one given lexeme and the lexemes themselves.

We preliminarily consider an inflectional paradigm canonical if it satisfies the following criteria given in Table 1 [24]. To these criteria we add the ones in Table 2 that further define canonical paradigm shape.[3] Deviation from these criteria leads to non-canonical paradigmatic properties.

		COMPARISON ACROSS CELLS OF A LEXEME	COMPARISON ACROSS LEXEMES
1	COMPOSITION/STRUCTURE	*same*	*same*
2	LEXICAL MATERIAL (≈ shape of stem)	*same*	*different*
3	INFLECTIONAL MATERIAL (≈ shape of inflection)	*different*	*same*
4	OUTCOME (≈ shape of inflected word)	*different*	*different*

Table 1. Criteria for Canonical Inflection according to Corbett in [24].

		CANONICAL INFLECTION
1	FEATURE EXPRESSION	*There is no "mismatch between form and function" [4].*
2	STEM	*Each lexeme has exactly one stem that combines with a series of exponents.*
3	COMPLETENESS	*There exists exactly one form corresponding to the expression of a specific morphosyntactic feature structure.*
4	INFLECTION CLASS	*All forms of a lexeme are built from one single inflection class.*

Table 2. Additional criteria for Canonical Inflection.

In this work, we present a representation of five types of non-canonical inflection phenomena, namely *suppletion, deponency, heteroclisis, defectiveness* and *overabundance* within the inferential realisational model for inflectional morphology developed by Walther in [67].[4] In section 5 we show the impact these phenomena can have on the complexity of morphological descriptions.

3.2 Stem alternations/Suppletion

Suppletion comes in two types: *stem suppletion* and *form suppletion* [19]. Stem suppletion occurs whenever, inside a paradigm, the forms' exponents remain regular, but their stems vary. This is for example the case for the French verb *aller* 'to come' which, according to most descriptions, shows as much as four

[3] Among the additional criteria, criterion 2 derives directly from criterion 2 in [24] and criterion 4 can be seen as derived from criterion 3 in [24].

[4] This model does not include a separate formalisation of *syncretism*. Syncretism is modeled as a combination of heteroclisis and deponency. For a complete discussion thereof, see [67].

different stems, *all-, v-, i-* and *aill-*. Form suppletion corresponds to cases where a whole form is inserted in a paradigm cell that should canonically be filled by a certain stem and the exponent corresponding to this specific cell. Form suppletion is described in [11] for the French verbe *être* 'to be' in the present indicative. For this verb, the 1^{st} person plural form *sommes*, for example, does not show the regular 1^{st} person plural exponent *-ons* that canonically appears with corresponding forms of other verbs (see Table 3).

	SINGULAR	PLURAL
P1	***suis***	***sommes***
P2	*es*	*êtes*
P3	*est*	***sont***

Table 3. Form suppletion in the present indicative paradigm of French *être* 'to be'

LEXEME	TRANSLATION	STEM1	STEM2
ÂRÂSTAN	'to adorn'	*ârâst*	*ârâ*
ÂMUXTAN	'to learn	*âmuxt*	*âmuz*
RAQSIDAN	'to dance	*raqsid*	*raqs*

Table 4. Persian present and past stems

Suppletion can be more or less transparent in the sense that it can be regularly associated with variation in the feature structure of a given word. Thus, Iranian languages such as Persian, show a stem alternation mainly related to tense-alternation: Persian uses a STEM1 for the present tenses and a STEM2 for the past tenses. The STEM2 is also used for the infinitives and the participle, while STEM1 serves as a stem for imperative forms.

According to traditional descriptions [29] Latin verbs also display three distinct stems that are linked to specific morphosyntactic features and subparts of the inflectional paradigms, namely present, past and supine: *amo* 'I love', *amāvī* 'I loved', *amātum* 'loved'. Yet the distribution of these stems does not follow strictly transparent feature-form associations. The third stem, for example, is associated with the passive past participle, but also with the active future participle and the finite passive perfective forms. There is no explicit morphosyntactic feature that appears to trigger the use of the third stem. Yet the distribution of the third stem is regular over all regular Latin verbs. Thus they are *morphomic* in the sense of Aronoff [2].

Moreover, suppletion can be more or less massive a phenomenon. While the Latin data only concerns three different stems, French verbs show stem suppletion that extends to twelve different stems [12]. Bonami and Boyé [12] show that there are up to twelve different feature combinations that can trigger stem suppletion. They call these twelve combinations *stem spaces*. The stems belonging to the stem spaces are linked through *stem dependency*.

Yet, among those languages for which stem selection seems to be an expression of morphosyntactic features, such as the Iranian languages, further irregularity can still occur. Thus, Sorani Kurdish displays specific stem selection irregularities: As Persian, Sorani Kurdish has distinct stems for present and past tense forms (respectively STEM1 and STEM2). Usually passive stems are built from STEM1; yet for some verbs, the passive uses STEM2, while for a third type of verbs a specific passive stem is required [47, 63, 66] (see Table 5). Such addi-

tional irregularities need to be captured before a corresponding morphological description's complexity can be measured.

Passive Stem Formation	lexeme	stem1	Present passive Stem	Translation
stem1	kuštin	*kuš*	*kuš–rê*	'to kill'
stem2	ûtin	*ǐě*	*ût–rê*	'to say'
stem2	bistin	*bîe*	*bist–rê*	'to hear'
stem1 minus endvowel	kirdin	*ke*	*k–rê*	'to do, 'to make'
stem1 minus endvowel	dan	*de*	*d–rê*	'to give'
other	xwardin	*xo*	*xû–rê*	'to eat'
other	girtin	*gir*	*gîr–rê*	'to take'

Table 5. Sorani Kurdish Irregular Passive Stems

3.3 Deponency

Croatian nouns sometimes use singular forms to express plural [3]. This "mismatch between form and function" is what, following Baerman [4], we name *deponency*. Nouns are inflected according to a number of different declension classes. Some classes that are relevant for our discussion are shown in Table 6: the nouns *dete* 'child' and *tele* 'calf' inflect according to the singular pattern of respectively the a-stem and i-stem inflection classes. Using a singular inflection to express plural results in this mismatch between form and function.

	(feminine) a-stem *žena* 'woman'		(feminine) i-stem *stvar* 'thing'	
	singular	plural	singular	plural
NOM	*žen-a*	*žen-e*	*stvar*	*stvar-i*
ACC	*žen-u*	*žen-e*	*stvar*	*stvar-i*
GEN	*žen-e*	*žen-a*	*stvar-i*	*stvar-i*
DAT	*žen-i*	*žen-ama*	*stvar-i*	*stvar-ima*
INS	*žen-om*	*žen-ama*	*stvar-i*	*stvar-im*

Table 6. Croatian noun declension

	neut. -et∼a-stem *dete* 'child'		neut. -et∼i-stem *tele* 'calf'	
	singular	plural	singular	plural
NOM	*dete*	*deca*	*tele*	*telad*
ACC	*dete*	*decu*	*tele*	*telad*
GEN	*deteta*	*dece*	*teleta*	*telad*
DAT	*detetu*	*deci*	*teletu*	*teladi(ma)*
INS	*detetom*	*decom*	*teletom*	*teladi(ma)*

Table 7. Croatian deponent noun declension

The Surrey Morphology Group has collected a whole range of data on deponency phenomena in a large database.[5] Even though we will see in Section 4 that our model would not retain all these examples as instances of deponency, this database constitutes an excellent general overview of deponency phenomena.

The most often discussed example of deponency probably are the Latin deponent verbs, where active meaning is considered to be conveyed through passive morphology [41, 36, 4, 25]. However, we shall give an alternate analysis of this particular data in Section 4, showing that these verbs actually are not instances of deponency but rather constitute a textbook example of *heteroclisis* [68].

[5] `http://www.smg.surrey.ac.uk/deponency`.

3.4 Heteroclisis

Heteroclisis refers to the phenomenon where a lexeme's paradigm is built out of (at least) two, otherwise separate, inflection classes.

Examples of heteroclisis are (some) Slovak animal nouns. Indeed, in Slovak, most masculine animal nouns are inflected as masculine animate nouns in the singular, whereas they may (and for some lexemes, must) inflect as masculine inanimate nouns in the plural (except in specific cases, such as personification, which triggers the animate inflection even for plural forms) [70]. Compare for example the inflection of *chlap* 'boy', *dub* 'oak' and *orol* 'eagle' in Table 8.[6]

	MASCULINE ANIMATE *chlap* 'boy'		MASCULINE INANIMATE *dub* 'oak'		MASCULINE HETEROCLITE *orol* 'eagle'	
	SINGULAR	PLURAL	SINGULAR	PLURAL	SINGULAR	PLURAL
NOM	*chlap*	*chlap-i*	*dub*	***dub-y***	*orol*	***orl-y***
GEN	*chlap-a*	*chlap-ov*	*dub-a*	*dub-ov*	*orl-a*	*orl-ov*
DAT	***chlap-ovi***	*chlap-om*	*dub-u*	*dub-om*	***orl-ovi***	*orl-om*
ACC	*chlap-a*	*chlap-ov*	*dub*	***dub-y***	*orl-a*	***orl-y***
LOC	***chlap-ovi***	*chlap-och*	*dub-e*	*dub-och*	***orl-ovi***	*orl-och*
INS	*chlap-om*	*chlap-mi*	*dub-om*	*dub-mi*	*orl-om*	*orl-ami*

Table 8. Heteroclisis in Slovak masculine animal names inflection

3.5 Defectiveness

Defectiveness [5] refers to lexemes which display empty (missing) cells in their paradigm. Sometimes languages contain lexemes for which expected forms are simply non-existing; native speakers are not capable of building the corresponding forms. Whenever such forms are needed, they must be conveyed through forms belonging to a synonymous lexeme. This is for example what we can observe with *activa tantum*: transitive verbs that do not possess passive forms and must therefore borrow the forms from synonyms. Examples thereof are the Latin verbs *facere* 'make' and *perdire* 'destroy' with no passive morphology in the present tense. The missing passive forms are supplied by (different) active verbs, namely *fieri* 'become' and *perire* 'perish' [41]. These supply verbs are not just passives for the former ones, but also normal intransitives. Hence, they canot be counted as part of the defective verbs' paradigms. They possess their own independent paradigm and constitute independent lexical entries. Another example are the nouns called *pluralia tantum* which only exist in the plural, cf. English *trousers*, French *vivres* 'food supplies' or Slovak *Vianoce* 'Christmas'.

[6] Both *chlap* and *dub* have a regular inflection: *chlap* belongs to the standard inflection class for masculine animate stems ending with a consonant, whereas *dub* belongs to the standard inflection class for masculine inanimate stems ending with what is called a hard or neutral consonant in the Slavic linguistic tradition.

3.6 Overabundance

The obvious counterpart to defectiveness is the concept of *overabundance*. Overabundance occurs when cells of a paradigm contain more than one form. The notion has been introduced by Thornton and is discussed in [64] for Italian. Canonical overabundance characterises the case where *cell mates* of one given cell compete, without any morphological feature permitting to choose one over the other. Table 9 shows examples thereof for Italian verbs.

	CELL-MATE 1	CELL-MATE 2
'languish' 3PL.PRS.SUBJ	*languano*	*languiscano*
'possess' 3PL.PRS.SUBJ	*possiedano*	*posseggano*
'possess' 3SG.PRS.SUBJ	*possieda*	*possegga*
'possess' 1SG.PRS.SUBJ	*possiedo*	*posseggo*

Table 9. Overabundance in Italian [64]

In French, an example is given by the verb *asseoir* 'to sit' that has two different forms in most cells as shown in Table 10.[7] All French verbs in *-ayer* also exhibit systematic overabundance (see Table 11). Indeed, for some cells, these verbs may use two competing stems (in *-ay-* and in *-ai-*) and therefore have two different inflected forms, morphologically equivalent (although semantic, pragmatic, sociolinguistic and other constraints may interfere).

IND.PRES	SINGULAR	PLURAL
P1	*assois* *assieds*	*assoyons* *asseyons*
P2	*assois* *assieds*	*assoyez* *asseyez*
P3	*assoit* *assied*	*assoient* *asseyent*

Table 10. Overabundance in French *asseoir* 'to sit'

IND.PRES	SINGULAR	PLURAL
P1	*balaye* *balaie*	*balayons*
P2	*balayes* *balaies*	*balayez*
P3	*balaye* *balaie*	*balayent* *balaient*

Table 11. Overabundance in French *balayer* 'to sweep'

4 A formal model for inflectional morphology

4.1 Defining the relevant notions

Since the non-canonical phenomena described in section 3 are precisely the irregularities that add complexity to the description of a lexeme's paradigm, we need a model capable of completely formalising the relevant irregularities. Only

[7] See for example [14] for a longer discussion thereof.

then can we use the formalised descriptions to measure their complexity with appropriate complexity metrics.

We use the formal inferential realisational model for inflectional morphology described in [67]. In this model a lexeme is considered w.r.t. its formal participation in the inflectional process. Thus, we do not consider any specific semantics or possible derivational properties. In other words, we are here interested in the behaviour of what Fradin and Kerleroux refer to as *inflectemes* [30], as opposed to lexemes, and for which a (very) simplified definition could be "a lexeme minus its semantic and argument-structural information".

This model represents an inflecteme $\mathfrak{I}$ as the association of five defining elements : (1) the set of morphosyntactic feature structures $\mathfrak{I}$ can express, (2) the lexeme's morphosyntactic category, (3) a *stem formation* rule, (4) an *inflection rule*, (5) a *transfer rule*.

Defectiveness and Overabundance In our model, categories are assigned to sets of inflectemes that canonically share sets of morpho-syntactic features. Belonging to a specific category creates morphological expectation in the sense of Brown *et al.* [20] as to which features should be realised by independent forms. If these expectations are not met by an inflecteme's forms this inflecteme is considered defective. Thus, defectiveness is defined for an inflecteme $\mathfrak{I}$ as the property of not fulfilling its category driven expectations: there is at least one morphosyntactic feature structure that should be expressed by the $\mathfrak{I}$'s categorie's members for which no form is produced for $\mathfrak{I}$.

Whenever more forms are generated than what is expected of a given inflecteme (given its membership of a certain category), this inflecteme is considered overabundant. Thus, defectiveness and overabundance occur whenever the inputs and outputs of an inflection rule f are not in a 1 to 1 correspondance.

Let us consider the French nominal inflecteme $\mathfrak{I}$ of *vivres* 'food supplies' as an example. Concerning the feature NUMBER, French nouns are expected to express the set of feature-value pairs {NUMBER *singular*, NUMBER *plural*}. However, *vivres* produces a realisation for the feature structure {NUMBER *plural*} only. It is hence defective.

Conversely, the Italian data in 9 shows instances of overabundance. For example, the inflecteme of *languire* is such that the realisation associated with the feature structure {NUMBER *plural*, PERSON *3*, TENSE *present*, MODE *subj*} produces two forms, *languano* and *languiscano*.

Stem selection and suppletion: *stem zones* The stem formation rule and the inflection rule are used for expressing the morphomic dimension of inflectional paradigms belonging to a given lexeme.

Hence, stem alternation in Latin can be represented through the existence of three different stem zones that are sets of cells in which the stem realisation rules associated with expressible morphosyntactic features always produce one type of stem, as shown in Tables 12 and 13 for the features listed in Table 14.

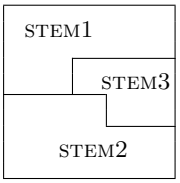

Table 12. A: Stem zones in the Latin active (sub-)paradigm

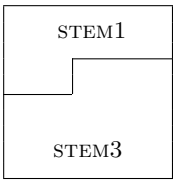

Table 13. B: Stem zones in the Latin passive (sub-)paradigm

STEM	ACTIVE SUBPARA.	PASSIVE SUBPARA.
STEM1	*imperf. finite*	*imperf.finite*
STEM2	*perf. finite*	
STEM3	*active future part.*	*passive past part.* *perf. finite (periphr.)*

Table 14. Morphomic feature association with Latin verb stems

Suppletion can hence be associated with specific stem zones. Moreover, the model allows for expressing that a given inflecteme $\mathfrak{I}$ is associated with specific stem zones through the notion of the inflecteme's *stem pattern*. In Table 14, the active subparadigm's stem pattern comprises STEM1, STEM2 and STEM3, while the passive subparadigm's stem pattern comprises only STEM1 and STEM3.

Form realisation: *inflection zones* In [67], an inflection class is defined as the default association of morpho-syntactic features with form realisation rules that apply to stem zones of a given inflecteme.

Just as we have defined stem zones, we then define an *inflection zone* as denoting the behaviour of a particular inflection class for a given set of cells. More precisely, each inflection class can be partitioned into inflection zones. In combination with stem zones, inflection zones allow for modeling situtations in which, for example, a given set of exponents is applied to two different stems of the same inflecteme for expressing different morphosyntactic features: the same inflection zone will thus be involved twice in the same paradigm. As sketched above and shown in more details in section 4.1, inflection zones and stem zones allow for a novel analysis [68] of so called Latin deponent verbs [41, 61, 36].

Deponency Another non-canonical phenomenon that may occur is *deponency*. As said above, we follow Baerman [4] in defining deponency as a "mismatch between form and function". This mismatch occurs whenever the features to be expressed by an inflecteme do not match the features usually expressed by a specifc realisation rule. This fact is captured by the notion of *transfer rule*, which takes as input a set of features to be expressed and outputs the set of features corresponding to the appropriate realisation rule. Canonically, the transfer rule is the identity function. An inflecteme is considered deponent whenever an inflecteme's transfer rule differs from the identity function.

In order to model the Croatian data from Table 6, we can thus use the transfer rule. Recall that Croatian sometimes uses singular forms to express plural [3]. In our model, an inflecteme $\mathfrak{I}$ functioning this way has a transfer rule $\mathscr{T}_{\mathfrak{I}}$ such that $\mathscr{T}_{\mathfrak{I}}(\{\text{NUMBER } \textit{plural}\})=\{\text{NUMBER } \textit{singular}\}$.

Heteroclisis Moreover, for Croatian deponent nouns, the inflection rule f outputs the zones in Table 15 for the irregular nouns in Table 7.[8] A, B and C correspond to the three different inflection classes illustrated in Table 7. The nouns *dete* 'child' and *tele* 'calf' use exponents from two different inflection classes each to buid their paradigm. It is thus heteroclite.

INFLECTION CLASS	A: NEUTER -ET-STEM	B: (FEMININE) A-STEM	C: (FEMININE) I-STEM
dete 'child'	SG: zone$_{A,sg}$	PL: zone$_{B,sg}$	
tele 'calf'	SG: zone$_{A,sg}$		PL: zone$_{C,sg}$

Table 15. Croatian noun inflection zones for deponent lexemes.

A similar analysis can be made of Latin "deponent verbs". Latin "deponent verbs" show morphological passive ("m-passive") forms, but express active syntax ("s-active"). Therefore, they are usually considered instances of deponency in the sense of [4]. On the basis that applying passive morphology to Latin verbs does not necessarily entail applying passive value, as shown in [41],[9] we consider that there are distinct inflection classes applying mainly to active vs. passive morphological forms ("m-passive"): changing a verb's inflection class is seen as a derivational process. Since there are distinct endings for m-active and m-passive, we claim that there must be distinct inflection rules, i.e., for every inflecteme, distinct pairings between specific morphosyntactic feature structures and inflection zones belonging to specific inflection classes, see Figures 16 and 17. Based on our definition of inflection zones, deponent verbs can be analysed as heteroclite[68], most of their endings being retrieved through inflection zones belonging to a Class B while the additional forms are retrieved from zones in a Class A (namely A3 for the active participles and A4 for the gerunds).

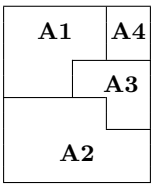

Table 16. Zones in Class A

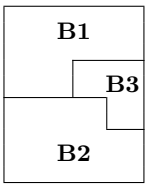

Table 17. Zones in Class B

LEXEME TYPE	M-ACTIVE	M-PASSIVE
ACTIVES	*A1, A2, A3, A4*	
PASSIVES		*B1, B2, B3*
DEPONENTS	*A3, A4*	*B1, B2, B3*
SEMI-DEP. T1	*A1, A3, A4*	*B2, B3*
SEMI-DEP. T2	*A2, A3, A4*	*B1, B3*

Table 18. Zone distribution in Latin verb inflection

[8] The representation shows that, in addition to being deponent, Croatian nouns are also heteroclite. Non-canonical behaviours can sometimes combine.

[9] Indeed, Kiparsky [41] shows that passive morphology can trigger many kinds of, partly unpredictable, semantic changes. This property is one of derivational morphology — and not inflectional morphology which is usually considered as being semantically predictable [17].

Given an inflecteme $\mathfrak{I}$, a pair formed by an inflection zone and a corresponding stem zone is called a *subpattern*. The complete *inflection pattern* of $\mathfrak{I}$, which consists of a set of subpatterns, allows for building all of $\mathfrak{I}$'s inflected forms. For example, the inflection of a passive Latin verb is fully defined by the following set of subpatterns: B1+STEM1, B2+STEM3, B3+STEM3. They constitute the inflection pattern of all such verbs.

Slovak animal nouns also show an instance of heteroclisis. As shown in Table 8, the zones for building the singular forms of the noun *orol* 'eagle' are partitions of the animate inflection class like those used for inflecting *chlap* 'boy', whereas the zones for the plural are retrieved in the inanimate inflection class, like for *dub* 'oak'.

Canonical Inflection It follows from the above-described irregularities that *Canonical inflection* corresponds to the case where

- an inflecteme's $\mathfrak{I}$ inflection pattern and stem pattern consist of only one (inflection *resp.* stem) zone each,
- the inflection rules produce exactly one possible realisation for each morphosyntactic feature structure expressable by $\mathfrak{I}$'s category,
- there is no mismatch between form and function, i.e. each exponent realised by a given realisation rule for a given morphosyntactic feature structure exactly corresponds to the morphosyntactic feature structure usually expressed in combination with this exponent.

5 Measuring the complexity of various descriptions of French verbal inflection

We have shown how non-canonical inflectional phenomena can be encoded in the model of inflectional morphology described in [67], using new notions such as inflection and stem zones. They can be viewed as generalisations of Bonami and Boyé's [12] stem spaces (and before them [52]), which, in turn, are correspond to stem pattern in this model. With such a formalism, various competing analyses for the same data can be designed, implemented, and therefore quantitatively evaluated with a suitable complexity measure. Not only does this provide a way to compare such analyses w.r.t. their complexity, but it is also a way to get insights into the relevance of these new notions, by examining whether they are used in analyses that have a lower complexity.

For answering these questions, we have developed and implemented a formalism capable of representing the model described in section 4. The basis for our formalism is the morphological layer of the Alexina lexical formalism [56] used by several morphological (and, for some, syntactic) lexica. We have extended this formalism in order to allow it to deal with inflection zones, transfer rules, patterns and stem patterns.

Next, we have encoded various competing morphological descriptions of French verbal inflection in this formalism, in order to assess the relevance of these newly

introduced notions and to quantatively compare these descriptions by means of the notion of *complexity*.

5.1 Descriptions of French verbal inflection

French verbal inflection is interesting in many well-known aspects, some of which have been described above. First, it is a rich system that generates forms corresponding to up to 40 different morphosyntactic features. Second, and this is of particular relevance when trying to assess the complexity of morphological descriptions, it is traditionally described as having one regular and productive inflection class, the class of so-called first-group verbs (verbs in *-er*), one irregular inflection class, that of third-group verbs, and the inflection class of second-group verbs (verbs like *finir*), which is sometimes considered as regular, as in traditional grammars, and sometimes as irregular. Analyses differ about the real productivity and regularity of this class [40, 18, 15], which is one first possible source of discrepancy between different accounts of French verbal inflection.

Among first-group verbs, as described in Section 3.6, verbs in *-ayer* exhibit (regular) overabundance. In [14], the authors consider them as polyparadigmatic. This is not fully satisfactory given the fact that both supposed paradigms would share the same forms for half of the cells. Another way to represent this situation is to define two stems, one in *-ay-* and one in *-ai-*, two inflection zones: one, ζ_1, that will be used only by the *-ay-* stem, and one, ζ_2, that will be used by both stems. Therefore, there would be three subpatterns within the (specific) inflection pattern for *-ayer* verbs: ζ_1+*-ay-*, ζ_2+*-ai-* and ζ_2+*-ay-*.

Modeling second-group verbs can also be achieved in different ways. Using Bonami and Boyé's [12] twelve-stem approach, these verbs can explicitely specify a secondary stem in *-iss* in the lexicon, along with the base *-i* stem (*fini-* vs. *finiss-* for *finir* 'finish'). The traditional (and widespread) way to represent this inflection class is to consider that it uses suffixes that begin in *-ss-* in certain cells. Obviously, this is not very satisfying. But as it happens, the cells for which second-group verbs use their secondary stem are exactly those which are covered by the zone ζ_1 defined in the previous paragraph. Therefore, if one defines a unique inflection class for first- and second-group verbs, the same ζ_1 and ζ_2 can be used here as well, together with the following stem pattern: starting from the base stem in *-i*, the secondary stem can be obtained through the addition of *ss*, while the inflection pattern is defined by two subpatterns, namely ζ_1+*-iss-* and ζ_2+*-i-*. Note that this corroborates the empirically grounded analysis in [65].

As for third-group verbs, the only two approaches that we have considered are the traditional one, using many inflection classes, and the twelve-stem approach by Bonami and Boyé [12]. Representing the latter approach in our model can be easily achieved, by modeling (default) stem dependencies within a stem pattern, and specifying for each verb (only) those stems that differ from what can be regularily obtained using the defined stem pattern.

Starting from these considerations, we have developed four different descriptions of French inflection in the new version of Alexina that implements our morphological model, in order to try and measure their respective complexity.

5.2 Quantifying and measuring morphological complexity[10]

In recent years, finding appropriate means to measure language complexity has become an active area of research. In increasing order of specificity, work has been done in that direction by considering languages globally [48, 38], by restricting the study to one particular level such as morphology [7], and by measuring the complexity of particular morphological descriptions, most notably in the context of unsupervised or weakly supervised learning of morphology [32, 69].

Various metrics for measuring linguistic complexity can be found in the litterature. The simplest ones simply count the occurrences of a handcrafted set of linguistic properties: size of various inventories (e.g., phonemes, categories, morph(eme) types...) [48]. However, such approaches are intrinsically arbitrary: both the set of properties which are chosen and the criteria underlying the way these properties are described are very hard to define in a principled way (e.g., what would be a suitable objective and language-independant way to build an inventory of categories for any language?). Alternative ways of measuring complexity rely on definitions of complexity that come from information-theoretic considerations. Two distinct definitions have been used in recent work, which apply on any kind of message and not only on linguistic descriptions or models: *information entropy* (or Shannon complexity), whose main drawback is that it requires encoding the message as a sequence of independent and identically-distributed random variables according to a certain probabilistic model, which is difficult in practice; and *algorithmic entropy* (or Kolmogorov complexity) which is a more general and objective measure of the amount of information in a message, but which is not directly computable and has to be approximated.

The Kolmogorov complexity, because it is more general, is more appealing. It relies on the following intuitive idea: a model is more complex than another if it requires a longer message to be described. However, since its computation is not directly possible, one often reduces the problem to computing some kind of entropy within a particular space of possible models, by using an approximation of the Kolmogorov complexity that is defined over this model space: the result is called the *description length* w.r.t. the model. This is the basis of the paradigm called *Minimum Description Length* [55]. Therefore, computing an approximation of the Kolmogorov complexity of a linguistic description requires to define as optimal as possible a way to encode this description as a string (the "code"), and then a means of computing an approximation of the Kolmogorov complexity of that (coded) string [7]. Moreover, a linguistic model is often structured, contrarily to what studies involving morphological complexity sometimes assume. In particular, assessing the complexity of a representation of a morphological lexicon cannot be reduced to measuring the complexity of a corpus whose forms have been segmented into morphs — which is however the basis of pioneering work in automatic acquisition of morphological information [32].

In our case, we want to measure the complexity of a given description of (a given part of) a morphological description of a particular language. This

[10] This title is borrowed from [7], whose first sections provide a brief but complete and detailed account of recent work on this topic.

is to be contrasted with cross-linguistic comparative studies on morphological (or linguistic) complexity in general [48, 38, 7]: we do not want to estimate the complexity of a language, but that of particular descriptions of its morphological component, and, more specifically, of its lexical inflectional system.

The description length $DL(m)$ of an unstructured message m within a model that decomposes it as a sequence of N symbols taken from an alphabet $W = \{w_1, \ldots, w_n\}$ can be computed as: $DL(m) = -\sum_i o(w_i) \log_2 o(w_i)/N$, where $o(w_i)$ is the number of occurrences of w_i in m. This description length is equal to N times the entropy of the message.

In our case, the code can not be that simple, as a morphological description is structured. First, as explained earlier, it is decomposed into the morphological lexicon and the morphological model. In our formalism, we define a lexical entry as a citation form, an inflection pattern, an optional non-default stem pattern and an optional list of non-predictable stems (predictable stems need not be specified). As for the morphological model, it involves patterns and subpatterns, tables, zones and form formation rules, sandhi rules[11] and other factorisation devices (see below for examples). We have designed a code that encodes all this structure in a bijective way (it can be non-ambiguously decoded) using symbols from 16 different alphabets (one for letters in citation forms, one for morphosyntactic tags, one for pattern ids, one for structural information within tables, and so on). As shown by preliminary experiments, the use of various alphabets leads to shorter descriptions as measured by the following generalisation of the above-mentioned formula: if a message m is decomposed loselessly in a sequence of symbols taken from the union of p alphabets $W^1, \ldots, W^p$, $W^1 = \{w_1^1, \ldots, w_{n_1}^1\}, \ldots, W^p = \{w_1^p, \ldots, w_{n_p}^p\}$ (i.e., the alphabet from which a given symbol is taken can be inferred deterministically from its left context), then we define its description length as:

$$DL(m) = -\sum_{j=1}^{p} \sum_{i=1}^{n_p} o(w_i^p) \log_2 \frac{o(w_i^p)}{N_p},$$

where N_p is the number of symbols from alphabet W^p in m. Such a metric allows for approximating the complexity of a structured model, and to measure the contribution of each alphabet to that complexity. This is the way we computed the complexity of various morphological descriptions of French verbal inflection in our model, both for evaluating the relevance of the newly introduced concepts (e.g., inflection zone, inflection pattern) and for comparing these competing morphological descriptions.

[11] We define *sandhi rules* as morphographemic and/or morphophonemic rules, already implemented in the Alexina formalism. They are local transformations that apply at the boundary between two morph(eme)s. Hence, in French verbal inflection, a stem ending in *-g* followed by a suffix in *[aou]-* is associated with a surface form in which an *e* is appended to the stem: *mang_ons* $\leftrightarrow$ *mange_ons*.

5.3 Measuring the complexity of various morphological descriptions of French verbal inflection

We have described above a spectrum of possible descriptions that correspond to various ways to balance richer morphological grammars and richer lexical specifications. We used the lexical information in the lexicon Le*ffff* [56] for our experiments, limiting ourselves to verbs and ignoring multiple entries for the same lexeme (a given lexeme may have several sub-categorisation patterns and several meanings, and therefore have several entries in the Le*ffff*). The current version of the Le*ffff* contains 7,820 verbs, among which 6,966 first-group verbs, 315 second-group verbs and 539 third-group verbs.

In our new version of Alexina's morphological layer, the morphological information associated with a lexical entry contains the following elements, illustrated by the example below:

- a citation form, typically the infinitive for French verbs;
- an inflection pattern followed by an optional pattern variant: if two patterns only differ on a few slots, they can be merged, and alternate realisation rules are specified for these slots and are lexically triggered by these infection pattern variants;
- optionally, a stem pattern (a non-specified stem pattern means that the default stem pattern associated with the pattern should be used);
- optionally, a list of stems (a non-specified stem means that the default stem should be used, as defined by stem formation rules associated with the stem pattern)

For example, an entry such as "bouillir $v_{23r}/_{bouill,,bou}$" corresponds to an inflecteme with the citation form *bouillir*, the pattern *v* (with the pattern variant *23r*), the default stem pattern associated with pattern *v* in the morphological grammar, as well as *bouill* as stem 1 and *bou* as stem 3 (all other stems following the stem pattern). Let us now briefly describe our four competing descriptions of French verbal inflection. The lexical entries for a small set of inflectemes in each of these descriptions is shown in Table 19 for illustration purposes.

At one end of the spectrum, we automatically generated a "flat" morphological description, called Flat, that uses no stems, no sandhi and no zones, in the following way. The longest common substring shared by all inflected forms of each lexeme has been identified, and the remainder of each form has been considered a "suffix"; then the list of all suffixes has been ordered w.r.t. the corresponding morphosyntactic tags, thus creating a *signature*. Finally, all lexemes that share the same signature have been considered as belonging to the same inflection class which is trivially built from the signature and the ordered list of tags. The resulting description has 139 inflection classes. Its description length, measured as explained above, is around 131,400 bits (9,200 bits in the lexicon,[12] 122,200 bits in the morphological grammar).

[12] Here and in all subsequent figures, the description length of citation forms is not taken into account, as it is the same for all descriptions.

Citation form	Flat	Orig	New	BoBo
aimer	v1	v-er_{std}	v-er	v_{1}
acheter	v18	v-er_{std}	v-er	v_{1}
jeter	v8	v-er_{dbl}	v-eter	$\text{v}_{1}/_{\text{,jett,,,,,,,,jettə}}$
balayer	v12	v-ayer	v-ayer	v-ayer_{1}
finir	v2	v-ir2	v-ir2	$\text{v}_{23\text{r}}$
requérir	v42	v-ir3	v-ir3	$\text{v}_{23\text{r}}/_{\text{requér,requier,,,,,,,,requer,requi,requis}}$
cueillir	v51	v-assaillir	$\text{v}_{23\text{r}}/_{\text{cueill,,,,,,,,,cueillə}}$	$\text{v}_{23\text{r}}/_{\text{cueill,,,,,,,,,cueillə}}$
prendre	v24	v-prendre	v-prendre	$\text{v}_{3\text{re}}$
mettre	v17	v-mettre	v-mettre	$\text{v}_{3\text{re}}/_{\text{,,met,,,,,,,,mi,mis}}$

Table 19. Lexical entries for a small set of inflectemes in each of our four competing descriptions of French verbal inflection

At the other end of the spectrum lies Bonami and Boyé's [12, 15] analysis, which uses only one inflection class and twelve stems. We started from a preliminary DATR implementation of this model (Bonami, p.c.). Because this analysis was designed on phonemes, we had to apply certain transformations to enable the encoding of graphemic inflection, including by introducing sandhi operations. In order to correctly generate all overabundant forms, we extended it in several ways. The result is a description, called BoBo, that contains only one inflection class, several patterns (4 for non-defective verbs including "v" and "v-ayer" found in Table 19, and a few more for defective ones) and 61 sandhi rules.[13] This description strongly relies on an important feature of our Alexina implementation of the model described in this paper and already mentioned above: any underspecified piece of information is filled by defaults (not specifying a given stem in a lexical entry leads to using the stem formation table for generating it; not specifying a stem generation table for a given lexical entry leads to using the default stem generation table associated with its inflection pattern, and, if not specified, to consider that there is only one stem that applies to all forms, and so on). BoBo's description length is around 52,000 bits (46,600 bits in the lexicon, which is particularly high and is caused by all explicitely specified stems, and 5,400 bits in the morphological grammar, which is very low as expected).

Between these two extremes, the original description Orig used by the L*efff*, which heavily relies on sandhis but uses a lot of inflection classes for third-group forms, has a description length of 83,000 bits (8,100 bits in the lexicon, 74,800 bits in the grammar). More interestingly, as mentioned above, using the notion of inflection zone and relying on a reasonable amount of sandhi rules, we were able to develop a more satisfactory morphological description for French verbs, named New, which uses 20 inflection classes (including one for first-group verbs without overabundance, one for first-group verbs in *-ayer* and one for second-group verbs). The corresponding description length, 35,800 bits, is lower than that of BoBo. It corresponds to 20,100 bits in the lexicon (twice more than in Flat, but twice less than in BoBo) and 15,700 bits in the morphological grammar (three times more than in BoBo, but eight times less than in Flat).

[13] For example, one of these rules handles the ə at the end of some of the stems, depending on its environment.

Description name	Flat	Orig	New	BoBo
Length of the morphological grammar (bits)	122,200	74,800	15,700	5,400
Length of the lexicon (bits)	9,200	8,100	20,100	46,600
Total length of the morphological description (bits)	131,400	83,000	35,800	52,000

Table 20. Description length of various accounts of French verbal morphology

All these figures for our four descriptions, ordered according to the above-mentioned continuum, are summarized in Table 20 and displayed graphically in Figure 1. They make it visible that using the notion of inflection zone, thus generalising the notion of stem space (which in our model corresponds to the notion of stem pattern), leads to accounts of French verbal morphology that constitute a shorter coding of the same information than the three other descriptions, both traditional ones and more original and recent ones [15]. Note that this conclusion would have been different if the description length of the lexicon had not been taken into account. However, as the balance between including more information in the lexicon and more in the morphological model depends on the morphological description, it would make little sense to evaluate the description length of the morphological model only.

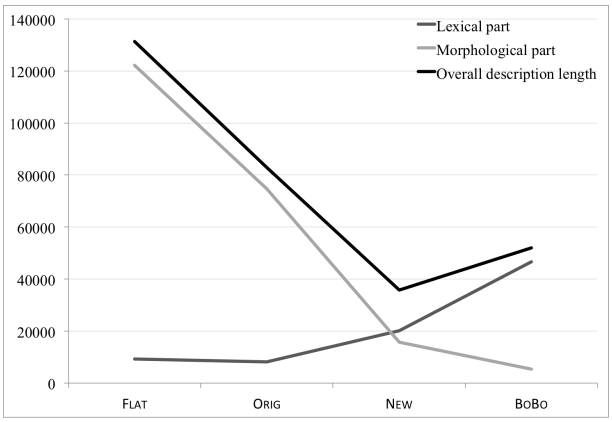

Fig. 1. Visualisation of the description length of various accounts of French verbal morphology

6 Conclusion

In this paper, we have addressed the question of measuring the complexity of morphological descriptions using the information-theoretic concept of description

length. We have applied our method on four competing descriptions of French verbal inflection.

Since descriptive complexity arises with inflectional irregularities, we have couched our descriptions in the formal inferential realisational model developed in [67]. This model, which relies on new notions such as the one of inflection zone and stem zone, allows for modeling a wide range of non-canonical inflectional phenomena, such as suppletion, deponency, heteroclisis, defectiveness and overabundance. We have developed four descriptions of French verbal inflection in this model and implemented them in the Alexina [56] morphological framework. We have also designed an information-theoretic way to assess the complexity of a morphological description in this model.

Our work shows that using information-theoretic concepts to assess description complexity is indeed feasable and relevant as a comparison between competing descriptions. Moreover, quantitative results on our four different descriptions have shown that the traditional way of describing French verbal inflection using many inflection classes, as well as a more recent and radically different proposal [15], can both be outperformed in terms of low complexity by using notions such as inflection zones, stem patterns and inflection patterns, in order to find a better balance between the amount of morphological information that is encoded in the lexicon and in the morphological rules.

References

1. Anderson, S.R.: A-morphous Morphology. Cambridge University Press, Cambridge, UK (1992)
2. Aronoff, M.: Morphology by Itself. MIT Press (1994)
3. Baerman, M.: Deponency in serbo-croatian. Online Database: `http://www.smg.surrey.ac.uk/deponency/Examples/Serbo-Croatian.htm` (2006)
4. Baerman, M.: Morphological Typology of Deponency. In: Baerman, M., Corbett, G.G., Brown, D., Hippisley, A. (eds.) Deponency and Morphological Mismatches. vol. 145, pp. 1–19. The British Academy, Oxford University Press (2007)
5. Baerman, M., Corbett, G.G., Brown, D.: Defective Paradigms: Missing forms and what they tell us. Oxford University Press, Oxford, UK (2010), proceedings of the British Academy 145
6. Baerman, M., Corbett, G.G., Brown, D., Hippisley, A. (eds.): Deponency and Morphological Mismatches. Oxford University Press (2007)
7. Bane, M.: Quantifying and measuring morphological complexity. In: Chang, C.B., Haynie, H.J. (eds.) Proceedings of the 26th West Coast Conference on Formal Linguistics. Sommerville, USA (2008)
8. Baroni, M., Matiasek, J., Trost, H.: Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In: Proceedings of the ACL Workshop on Morphological and Phonological Learning. pp. 48–57 (2002)
9. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI (2003)
10. Bernhard, D.: Apprentissage non supervisée de familles morphologiques par classification ascendante hiérarchique. In: Proceedings of TALN 2007. pp. 367–376. Toulouse, France (2007)

11. Bonami, O., Boyé, G.: Suppletion and dependency in inflectional morphology. In: Eynde, F.V., Hellan, L., Beerman, D. (eds.) Proceedings of the HPSG '01 Conference. CSLI Publications, Stanford, USA (2002)
12. Bonami, O., Boyé, G.: Supplétion et classes flexionnelles dans la conjugaison du français. Langages 152, 102–126 (2003)
13. Bonami, O., Boyé, G.: Deriving inflectional irregularity. In: Proceedings of the 13th International Connference on HPSG. pp. 39–59. CSLI Publications, Stanford, USA (2006)
14. Bonami, O., Boyé, G.: La morphologie flexionnelle est-elle une fonction? In: Choi-Jonin, I., Duval, M., Soutet, O. (eds.) Typologie et comparatisme. Hommages offerts à Alain Lemaréchal, pp. 21–35. Peeters, Leuven, Belgium (2010)
15. Bonami, O., Boyé, G., Giraudo, H., Voga, M.: Quels verbes sont réguliers en français? In: Actes du premier Congrès Mondial de Linguistique Française. pp. 1511–1523 (2008)
16. van den Bosch, A., Daelemans, W.: Memory-based morphological analysis. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics. pp. 285–292 (1999)
17. Boyé, G.: Régularité et classes flexionnelles dans la conjugaison du français. In: Roché, M., Boyé, G., Hathout, N., Lignon, S., Plénat, M. (eds.) Des unités morphologiques au lexique. Hermes Science (2011)
18. Boyé, G.: Problèmes de morpho-phonologie verbale en français, espagnol et italien. Ph.D. thesis, Université Paris 7 (2000)
19. Boyé, G.: Suppletion. In: Brown, K. (ed.) Encyclopedia of Language and Linguistics (2nd ed.), vol. 12, pp. 297–299. Elsevier, Oxford, UK (2006)
20. Brown, D., Chumakina, M., Corbett, G.G., Popova, G., Spencer, A.: Defining 'periphrasis': key notions (2011), under editorial review.
21. Cartoni, B.: Lexical morphology in machine translation : A feasibility study. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). pp. 130–138. Athens, Greece (March 2009)
22. Chomsky, N., Halle, M.: The sound pattern of English. Harper and Row (1968)
23. Corbett, G.G.: Agreement: the range of the phenomenon and the principles of the surrey database of agreement. Transactions of the philological society 101, 155–202 (2003)
24. Corbett, G.G.: Canonical typology, suppletion and possible words. Language 83, 8–42 (2007)
25. Corbett, G.G.: Deponency, Syncretism, and What Lies Between. In: Baerman, M., Corbett, G.G., Brown, D., Hippisley, A. (eds.) Deponency and Morphological Mismatches. vol. 145, pp. 21–43. The British Academy, Oxford University Press (2007)
26. Corbett, G.G., Fraser, N.: Network Morphology: a DATR account of Russian nominal inflection. Journal of Linguistics 29, 113–142 (1993)
27. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02. pp. 21–30 (2002)
28. Demberg, V.: A language-independent unsupervised model for morphological segmentation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 920–927. Prague, Czech Republic (June 2007)
29. Ernout, A., Thomas, F.: Syntaxe Latine. Kliensieck, Paris, 2 edn. (1953)
30. Fradin, B., Kerleroux, F.: Troubles with lexemes. In: Booij, G., Janet de Cesaris, Sergio Scalise, A.R. (eds.) Selected papers from the Third Mediterranean Mor-

phology Meeting. pp. 177–196. Topics in Morphology, IULA-Universitat Pompeu Fabra, Barcelona, Spain (2003)
31. Gaussier, E.: Unsupervised learning of derivational morphology from inflectional lexicons. In: Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing. University of Maryland (1999)
32. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics 27(2), 153–198 (2001)
33. Halle, M., Marantz, A.: Distributed morphology and the pieces of inflection. In: Hale, K., Keyser, S.J. (eds.) The view from building 20, pp. 111–176. MIT Press, Cambridge, USA (1993)
34. Harris, Z.S.: From phoneme to morpheme. Language 31(2), 190–222 (1955)
35. Hathout, N.: From wordnet to celex : acquiring morphological links from dictionaries of synonyms. In: Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 1478–1484. Las Palmas de Gran Canaria, Spain (2002)
36. Hippisley, A.: Declarative Deponency: A Network Morphology Account of Morphological Mismatches. In: Baerman, M., Corbett, G.G., Brown, D., Hippisley, A. (eds.) Deponency and Morphological Mismatches. vol. 145, pp. 145–173. The British Academy, Oxford University Press (2007)
37. Jacquemin, C.: Guessing morphology from terms and corpora. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 156 – 165 (1997)
38. Juola, P.: Assessing linguistic complexity. In: Miestamo, M., Sinnemäki, K., Karlsson, F. (eds.) Language Complexity: Typology, Contact, Change. John Benjamins Press, Amsterdam, Netherlands (2008)
39. Keshava, S.: A simpler, intuitive approach to morpheme induction. In: In PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes. pp. 31–35 (2006)
40. Kilani-Schoch, M., Dressler, W.U.: Morphologie naturelle et flexion du verbe français. Gunter Narr Verlag, Tübingen, Germany (2005)
41. Kiparsky, P.: Blocking and periphrasis in inflectional paradigms. In: Yearbook of Morphology 2004. pp. 113–135. Springer, Dordrecht, Netherlands (2005)
42. Koskenniemi, K.: A general computational model for word-form recognition and production. In: Proceedings of the 22nd annual meeting of the Association for Computational Linguistics. pp. 178–181 (1984)
43. Lepage, Y.: Solving analogies on words : an algorithm. In: Proceedings of the 17th international conference on Computational Linguistics. pp. 728–734 (1998)
44. Lieber, R.: Deconstructing Morphology : Word Formation in Syntactic Theory. University of Chicago Press, Chicago, USA (1992)
45. Lovis, C., Michel, P.A., Baud, R., Scherrer, J.R.: Word segmentation processing: A way to exponentially extend medical dictionaries. In: Greenes, R.A., Peterson, H.E., Protti, D.J. (eds.) Proceedings of the 8th World Congress on Medical Informatics. pp. 28–32 (1995)
46. Matthews, P.H.: Morphology. Cambridge University Press, Cambridge, UK (1974)
47. McCarus, E.N.: A Kurdish Grammar: descriptive analysis of the Kurdish of Sulaimaniya, Iraq. Ph.D. thesis, American Council of Learned Societies, New-York, USA (1958)
48. McWhorter, J.: The world's simplest grammars are creole grammars. Linguistic Typology 5, 125–166 (2001)

49. Moreau, F., Claveau, V., Sébillot, P.: Automatic morphological query expansion using analogy-based machine learning. In: Proceedings of the European Conference on Information Retrieval, ECIR 07. Rome, Italy (April 2007)
50. Namer, F.: Morphologie, lexique et tal : l'analyseur dérif. In: TIC et Sciences cognitives. Hermes Sciences Publishing, London (2009)
51. Pinker, S.: Words and Rules. Basic Books, New-York, NY, USA (1999)
52. Pirelli, V., Battista, M.: The paradigmatic dimension of stem allomorphy in italian verb inflection. Italian Journal of Linguistics pp. 307–380 (2000)
53. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
54. Pratt, A.W., Pacak, M.G.: Automated processing of medical english. In: Proceedings of the 1969 conference on Computational linguistics. pp. 1–23 (1969)
55. Rissanen, J.: Universal coding, information, prediction, and estimation. IEEE Transactions on Information Theory 30(4), 629–636 (1984)
56. Sagot, B.: The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In: Proceedings of the 7th Language Resource and Evaluation Conference. La Valette, Malte (2010)
57. Snyder, B., Barzilay, R.: Unsupervised multilingual learning for morphological segmentation. In: Proceedings of ACL-08. pp. 737–745. Columbus, USA (June 2008)
58. Spiegler, S., Golenia, B., Flach, P.: Promodes : A probabilistic generative model for word decomposition. In: Working Notes for the CLEF 2009 Workshop. Corfu, Greece (2009)
59. Stroppa, N., Yvon, F.: Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie. Traitement Automatique des Langues 47, 33–59 (2006)
60. Stump, G.T.: Inflectional Morphology. A Theory of Paradigm Structure. Cambridge University Press, Cambridge, UK (2001)
61. Stump, G.T.: Heteroclisis and paradigm linkage. Language 82, 279–322 (2006)
62. Tepper, M., Xia, F.: Inducing morphemes using light knowledge. Journal of ACM Transactions on Asian Language Information Processing (TALIP) 9(3), 1–38 (2010)
63. Thackston, W.: Sorani Kurdish: A reference grammar with selected readings (2006), `http://www.fas.harvard.edu/~iranian/{S}orani/sorani_1_grammar.pdf`, published online
64. Thornton, A.M.: Towards a typology of overabundance (December 2010), presented at the Décembrettes 7, Toulouse, France.
65. Tribout, D.: Les conversions de nom à verbe et de verbe à nom en français. Ph.D. thesis, Université Paris Diderot – Paris 7 (2010)
66. Walther, G.: A derivational account for Sorani Kurdish passives. (2011), presentation at the 4th International Conference on Iranian Linguistics (ICIL4). June 17th-19th 2011. Uppsala, Sweden
67. Walther, G.: PARSIL: an inferential-realisational model of inflectional morphology for the Paradigm Shape Lexicon Interface. Linguistica 51 (2011), internal and External Boundaries of Morphology. Accepted
68. Walther, G.: Latin passive morphology revisited (2011), presentation at the 2011 Meeting of the Linguistic Association of Great-Britain (LAGB 2011). September 7th-10th 2011. Manchester, UK
69. Xanthos, A.: Apprentissage automatique de la morphologie — Le cas des structures racine-schème, Sciences pour la Communication, vol. 48. Peter Lang (2008)
70. Zauner, A.: Praktická príručka slovenského pravopisu. Vydavateľstvo Osveta, Martin, Slovakia (1973)
71. Zweigenbaum, P., Grabar, N.: Liens morphologiques et structuration de terminologie. In: Actes de IC 2000 : Ingénierie des Connaissances. pp. 325–334 (2000)

72. Zwicky, A.M.: How to describe inflection. In: Niepokuj, M., Clay, M.V., Nikiforidiou, V., Feder, D. (eds.) Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society. pp. 372–386. Berkeley Linguistics Society (1985)