An open-ended computational construction grammar for Spanish verb conjugation

Beuls, Katrien

Published in:
Constructions and Frames

DOI:
10.1075/cf.00005.beu

Publication date:
2017

Document Version
Peer reviewed version

This paper is the authors' draft and has now been officially published as:

Beuls, K. (2017). An open-ended computational construction grammar for Spanish verb conjugation. Constructions and Frames 9(2): 278-301.

An Open-ended Computational Construction Grammar for Spanish verb conjugation

Katrien Beuls

Abstract

The Spanish verb phrase can take on many forms, depending on the temporal, aspectual and modal interpretation that a speaker wants to convey. At least half a dozen constructions work together to build or analyse even the simplest verb form such as *hablo* 'I speak'. This paper documents how the complete Spanish verb conjugation system can be operationalised in a computational construction grammar formalism, namely Fluid Construction Grammar. Moreover, it shows how one can, starting from a seed grammar handling regular morphology and grammar, create a productive grammar that can capture systematicity in Spanish verb conjugation that can expand its construction inventory when new verbs are encountered.

1 Introduction

The acquisition of verbal inflections is a crucial aspect in the process of gaining full expressivity and control over the situation and the internal conceptualization of events one is referring to. Inflected verbs carry information on tense, aspect and mood of the event, along with agreement information on the person and number of their subject, occasionally including agreement with the object of the event encoded by the verb (as in Hungarian, see for instance [Beuls, 2011]). The constructions involved in building such verb forms cut through multiple levels of linguistic knowledge, combining information from semantics, syntax, morphology and phonology. Romance languages are notorious for their extensive verbal paradigms that learners have to acquire on the path towards a productive command of a language such as French, Italian, Spanish or Portuguese. The current paper focuses on Spanish, the second largest language spoken across the globe after Mandarin and a very popular language when it comes to second language acquisition.

Mastering Spanish verbal inflections requires knowledge of both formal and semantic constraints that informs the learner of when to use which form. Formal constraints concern the usage of the appropriate stem of a verb lemma in a given slot of the verbal paradigm, with up to 120 different slots per verb stem and up to six allo-morphic stems per lemma. Moreover, phonological processes such as assimilation between stem and ending and word stress-induced stem changes increase the level of formal complexity of the verbal system. Meaning constraints that often take a long time to be assimilated by non-native speakers include the aspectual distinction between the past imperfect and the preterite use of an event (*cantaba* vs. *cant*, 'he sang' (imperfective-perfective)) as well as the semantic difference between the present perfect and the preterite (*ha cantado* vs. *cantó*).

The current contribution proposes a computational construction grammar account that can comprehend and produce any verb form in Spanish, including neologisms that obey the rules of Spanish word formation. This article explores target-language-tailored learning operators that assist in expanding an initial hand-coded "seed" grammar that contains the basic concatenative morphology constructions and verb phrase constructions needed to ensure an accurate conjugation of Spanish verbs. This grammar has been

designed originally in the setting of an intelligent tutoring system specialized in assisting the learner in acquiring this intricate system of forms and usages [Beuls, 2013]. Designing a grammar for real-life purposes requires the necessary robustness and extendibility to incorporate a large number of verbs currently in use in the Spanish language as well as to allow the occasional introduction of a new verb into the language by mechanisms of creative language use (e.g. *facebookear* 'to be active on Facebook') or the incorporation of foreign verbs into the grammar (e.g. *textear* 'to send text messages', alternative for *mandar mensajes*).

Allowing such innovations to extend and/or modify the construction inventory can be achieved in a number of ways in Fluid Construction grammar, and I will refer to these different options as degrees of productivity. First, the flexible matching of new verb forms when they are used in combination with endings that are part of the grammar opens the door for new lemmas to get introduced into the grammar. Yet, to productively conjugate a new lemma, its conjugation paradigm (verb class, stress pattern, stem changes) needs to be learned and incorporated into the grammar. Second, in transparent verbs, i.e. verbs that do not show a clear segmentation into stem and ending(s), can be acquired as they are encountered and saved as holistic constructions. Future usage of the same verb can then lead to the gradual discovery of a rule that supports its conjugation, thereby adding constructions needed to support this. Moreover, highly frequent verb forms are typically thought to be saved as a whole in the construction inventory of a speaker to speed up processing as the form does not have to be analyzed or constructed by several constructions [Croft and Cruse, 2004].

The article is structured as follows: Section 2 highlights the main learning challenges involved in the acquisition of Spanish verb conjugation for second language learners. The main constructions that are at work in the conjugation of Spanish verbs are described in Section 3. How such a seed grammar containing the basic semantic distinctions of the Spanish verbal system as well as the morphemes that make up its regular suffixes can be turned into a productive grammar by means of specific learning operators is explained in Section 4. Finally, Section 5 puts the findings of the current contribution into a larger perspective and includes hints for future explorations of this topic.

2 Learning challenges in the acquisition of Spanish verbs

Two main learning challenges are related to the acquisition of Spanish verbal morphology, which also require special attention in the formalization of this domain in Fluid Construction Grammar. The first challenge concerns semantic conceptualizations needed to realize a particular verb form. The second challenge is related to the range of different stems associated with a vast number of highly frequent verb forms.

2.1 Semantic challenges

Hypothesized to be perhaps the most difficult aspect in the acquisition of Spanish as a second language learner are the conceptualizations underlying the use of aspectual distinctions in the past tense domain (preterit vs. imperfective) as well as those present in the use of the subjunctive mood [Delbecque et al., 2001]. The following subsections treat the semantic space needed to express distinctions related to time, aspect and mood.

2.2 Tense

Tense is a grammatical category that reflects where an event is situated on the time line, with respect to the moment of speaking. A formal representation of the conceptualization

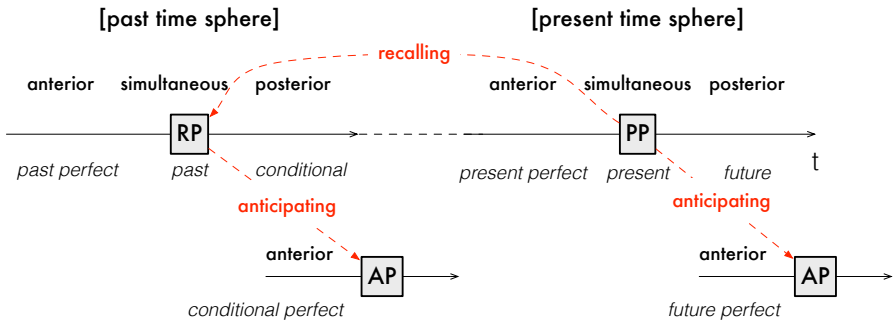


Figure 1: Spanish (like most Indo-European languages) makes use of two main time spheres represented by its systemic tense system: the past time sphere and the present time sphere. The Present Point (PP) is the anchor point in the present time sphere and corresponds to the utterance time. The Recalled Point (RP) is a point in the past recalled at the PP. Both from the PP and the RP an additional anchor point can be anticipated relative to which a situation is located (Anticipated Point). This figure has been adapted from (Bull, 1965, p. 113).

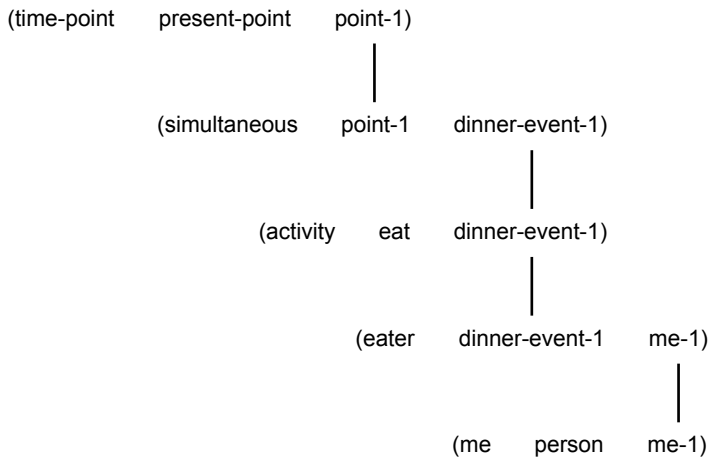
underlying the basic tense system in Spanish has been proposed by [Bull, 1965]. This standard formal account relies on two time axes: the present time sphere and the past time sphere. In Bull's model all situations are directly or indirectly related to the present point ('now'), which has also been referred to as a temporal zero-point (t_0) [Michaelis, 2006]. The time of utterance always functions as t_0. The tense system is then divided into two time-spheres: the past time-sphere and the present time-sphere. The past time-sphere is situated completely before t_0. The present time-sphere includes t_0 and is divided by it into three parts:

1. The pre-present sector: the part of the present time-sphere lying before t_0 (e.g. *he cenado*, 'I have had dinner'));
2. The present sector: the part of the present time-sphere centered around t_0 (e.g. *ceno*, 'I am having dinner');
3. The post-present sector: the part of the present time-sphere following t_0 (e.g. *cenaré*, 'I will have dinner').

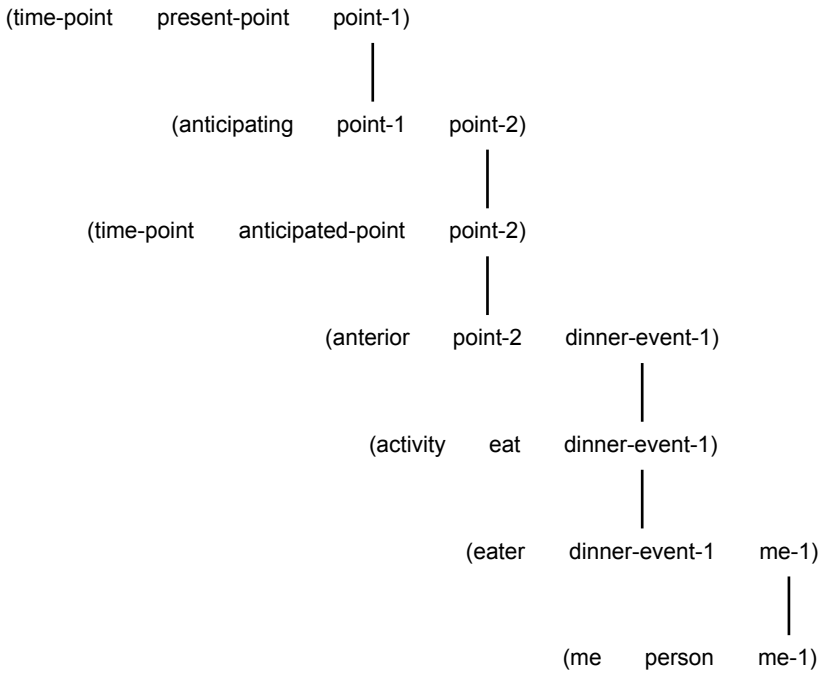
Absolute tenses like the past tense are typically described as exhibiting a direct relationship between utterance time and the time of the situation. Relative tenses such as the future perfect (e.g. *habré cenado*, 'I will have had dinner (by the time you come back).') express an indirect relationship in which the eating event is represented as the past relative to a point that is in the future relative to utterance time (coming home). The conditional perfect is another relative tense in which a point in time is anticipated from a recalled point such as in *habría cenado*, 'I would have had dinner'.

Figure 1 includes eight basic tenses as they are used in contemporary Spanish. Four of them are absolute tenses: the three tenses in the present time sphere directly related to the Present Point (PP, or t_0) and the past tense that is recalled directly from the PP, thereby establishing an anchor point in the past time sphere: Recalled Point (RP). The systemic meanings of the conditional and the past perfect tense can be conceptualized relative to this RP . Finally, the future and conditional perfective are represented as anterior to an Anticipated Point (AP), anticipated either from the PP or the RP (in the latter case establishing a double relative tense).

To operationalise this theory in a computational construction grammar account, the schema in Figure 1 was turned into corresponding meaning representations. An example of a present tense verb form meaning representation for *ceno* 'I have dinner' would be:



An example semantic representation of a relative tense such as the future perfect in *habré cenado* 'I will have had dinner' looks as follows:



2.3 Aspect

Aspect is a way of "viewing the internal temporal constituency of a situation" [Comrie, 1976] in terms of beginning, middle and end. Overt aspect marking happens through morphemes: e.g. *cant-abas* (imperfect) vs. *cant-aste* (preterite) 'you sang'. The preterite/imperfect distinction in Spanish morphology is one of the most debated topics in Spanish linguistics. Many thorough analyses have been proposed in the literature [Bull, 1965, Gili y Gaya, 1943, Ramsey, 1956] but none has captured it in a way that accounts for all cases supported by native speakers' intuitions. The imperfect is typically conceived as a means of bringing the listener to some recalled point (RP) in the middle of an event (or a series of events). The preterite can express "an occurrence from the viewpoint of either RP or PP, and it handles any aspect but middleness" [Whitley, 2002, p.117]. The semantic representations reflect the distinctions between imperfect and preterite in Spanish by means of a single meaning predicate `event-perspective`, with two values: `bound` for preterite events and `unbound` for imperfect events. Figure 2 shows the distinction in meaning representations between *cené* and *cenaba* 'I had dinner'.

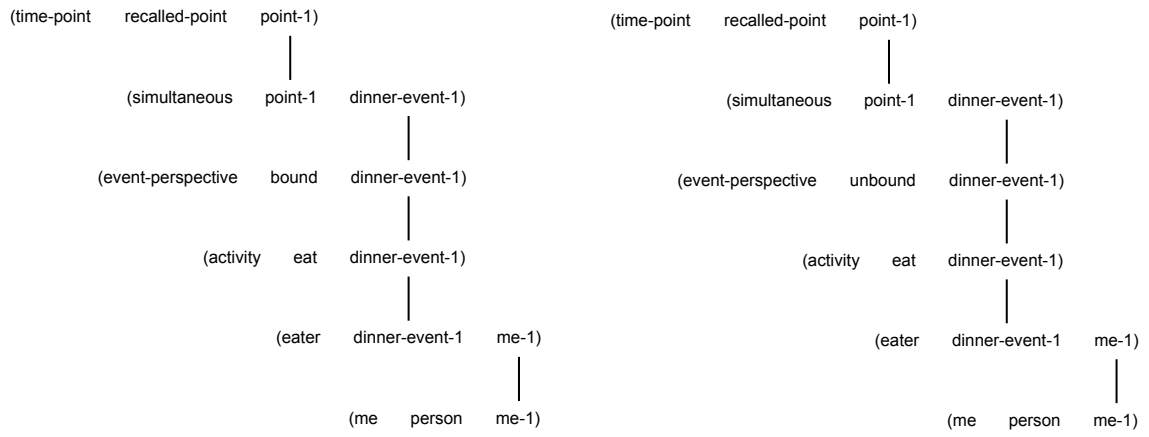


Figure 2: Different event perspectives distinguish between the meaning representations underlying *cené* (left) and *cenaba* (right). The preterite aspect describes how an event ended or began (therefore focusing on the boundaries), whereas the imperfect aspect highlights the ongoing nature of an event ('in the middle of').

2.4 Mood

The tense system explained in 2.2 applied to the indicative mood system. Spanish also uses a second system of tenses referred to as the subjunctive. Mood is a grammatical category that expresses the approaches a speaker can take with respect to a proposition. According to [Terrell and Hooper, 1974], there are six such approaches:

1. Asserted:
   (a) by the speaker: *me parece que* 'it seems that', *sé que* 'I know that', *es cierto que* 'it is certain that';
   (b) by others: *cuentan que* 'they say that', *se cree que* 'It is believed that';
2. Presupposed:
   (a) in the learning of a proposition: *se da cuenta de que* 'one realises that';
   (b) for commentary: *me alegro de que* 'I am happy that', *es lástima que* 'it is a pity that', *es interesante que* 'It is interesting that';
3. Neither:
   (a) its doubted: *duda que* 'doubts that', *niega que* 'negates that', *no creo que* 'I do not think that';
   (b) its willed: *manda que* 'mandates that', *quiere que* 'wishes that', *pido que* 'askes that'.

Typically, the indicative covers the first three cases (1a-2a), whereas the subjunctive is used in the last three (2b-3b). The difference between *dice que viene* (indicative) and *dice que venga* (subjunctive) can be translated into (1b) as opposed to (3b) [Whitley, 2002, p.131].

2.5 Morpho-syntactic challenges

Spanish verb endings are complex feature bundles with five main dimensions: person, number, tense, mood and aspect. Their richness is partly due to the absence of pronominal subjects in standard Spanish utterances such as habla 'he/she speaks', which implies

Table 1: The Spanish verb conjugation paradigm for all non-composed verb forms of a single verb mount to 59 forms. Minor levels of syncretism can be observed in forms such as *cenamos*, *cenaba*, *cené*, *cenaré* and *cenara/-se*. When composed forms are added, the number of forms doubles (118). Also politeness forms have been omitted from this table.

Impersonal forms			Imperative	
Infinitive	**Participle**	**Gerund**	**Singular**	**Plural**
cenar	cenado	cenando	cena	cenad
Indicative			Subjunctive	
Present	**Future**	**Conditional**	**Present**	**Future**
ceno	cenaré	cenaría	cene	cenare
cenas	cenarás	cenarías	cenes	cenares
cena	cenará	cenaría	cene	cenare
cenamos	cenaremos	cenaríamos	cenemos	cenáremos
cenáis	cenaréis	cenaríais	cenéis	cenareis
cenan	cenarán	cenarían	cenen	cenaren
Past imperfect	**Past perfect**		**Past imperfect**	
cenaba	cené		cenara/cenase	
cenabas	cenaste		cenaras/cenases	
cenaba	cenó		cenara/cenase	
cenábamos	cenamos		cenarámos/cenásemos	
cenabais	cenasteis		cenarais/cenaseis	
cenaban	cenaron		cenaran/cenasen	

that grammatical person and number information needs to be encoded in the verb forms themselves. Pronominal subjects are only expressed when they receive special emphasis in contrastive situations or focus expressions (*él cena* 'he has dinner'). The high amount of information encoded in Spanish verb forms results in a single lexeme having up to 120 different verb forms when its full conjugational paradigm is taken into account: 16 tenses/moods (8 tenses per mood, see above), seven inflected forms per tense/mood (including the politeness form), two infinitives, two gerunds and four participle forms [Bosque and Demonte, 1999]. Table 1 includes all non-composed forms in the conjugation of the regular verb cenar 'to have dinner'.

Apart from its three main verb classes (similar to Latin), Spanish verbs can be divided into three main groups of regularity patterns: regular verbs, irregular verbs and semi-regular verbs [Mayol, 2003]. Semi-regular verbs show regular patterns in most parts of their conjugation paradigm but are characterized by irregular forms in certain slots, mainly due to changing stress patterns and assimilation processes. Such semi-regular stems show a variety of stem realisations, e.g. the verb *tener* 'to have' has four allomorphic stems along with the default *ten-* stem: *tien-* (diphthongization), *teng-* (velar insertion), *tend-* (assimilation with following future tense *-r*) and the irregular *tuv-*. An introductory Spanish grammar identifies not less than 89 different conjugation schemes to cover the conjugation of all Spanish verbs [Mateo, 1998].

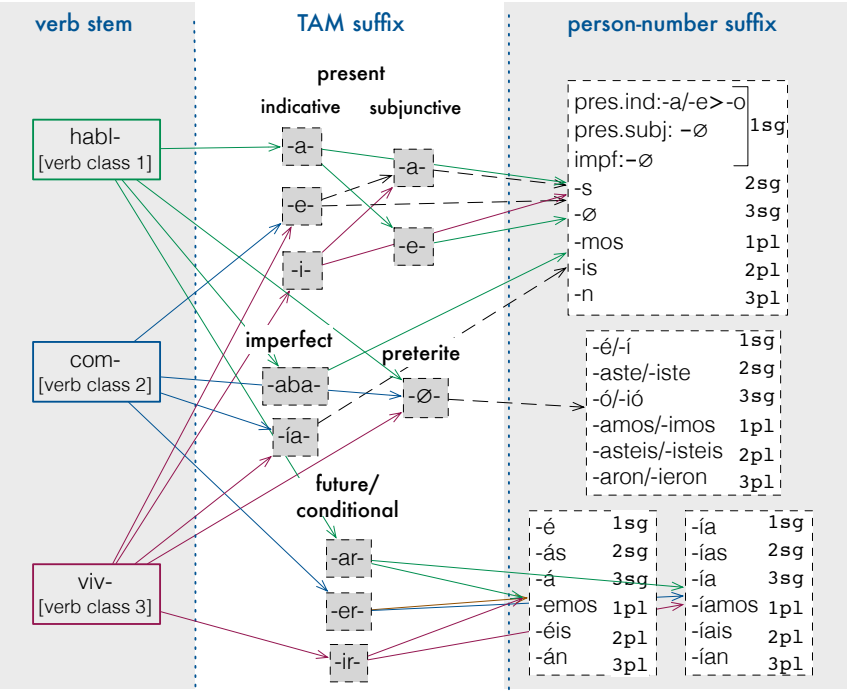


Figure 3: Non-periphrastic verb forms are combined out of three main building blocks: a verb stem, a tense-aspect-mood suffix and a person-number suffix. Due to the existence of three verb classes in Spanish, the tense-aspect-mood suffixes reflect the theme vowels of the verb's infinitive (*-ar*, *-er* or *-ir*). The subjunctive tam-suffixes are obtained by transforming the indicative tam-suffixes, a process in which both the *-e-* and the *-i-* collapse into *-a-*.

3 Spanish grammar fragment

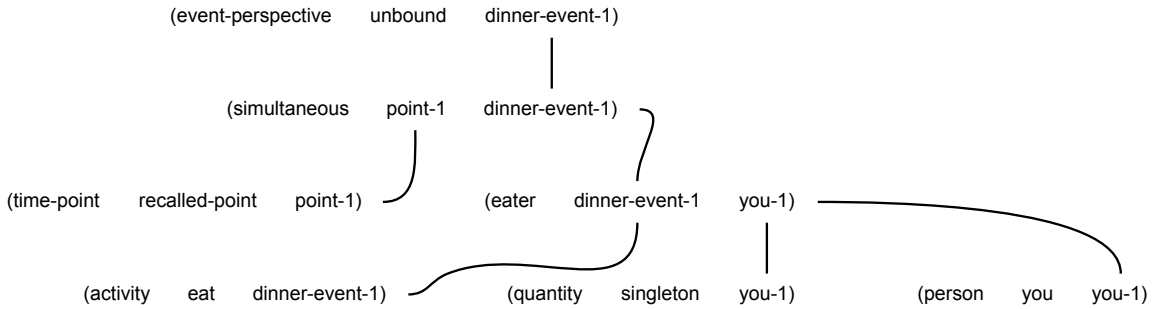
How are these linguistic challenges translated into an implementation in Fluid Construction Grammar? The current section first describes the grammar design, explaining the segmentation of verb forms and the types of constructions that were used, together with detailed explanations on their design (features in conditional and contributing locks). We then turn to the usage of these constructions to conjugate Spanish verb forms in production and comprehension in Section 3.2. Finally, Section 3.3 shows how the grammar fragment handles with allomorphic stem realizations. All examples can be tested in the online web demonstration that accompanies the current paper at `http://www.fcg-net.org/demos/spanish-verb-conjugation`.

3.1 Grammar design

With the main goal to create a productive grammar that can capture generalizations in the Spanish verb paradigm, verb forms had to be segmented into a stem and two suffixes. The first suffix indicates the tense, aspect, mood situation of the verb form and its form depends on the verb class of the stem. The second suffix expresses person and number information. A verb form such as *cantábamos* 'we sang' is thus segmented into the following three blocks: (1) its stem *cant-*, (2) a tense-aspect-mood suffix *-aba* and (3) a person-number suffix *-mos*. Each of these three blocks can be exchanged to create a new verb form. Impersonal forms such as the participle only have a single suffix *-ado* (or *-ido*) following the verb stem.

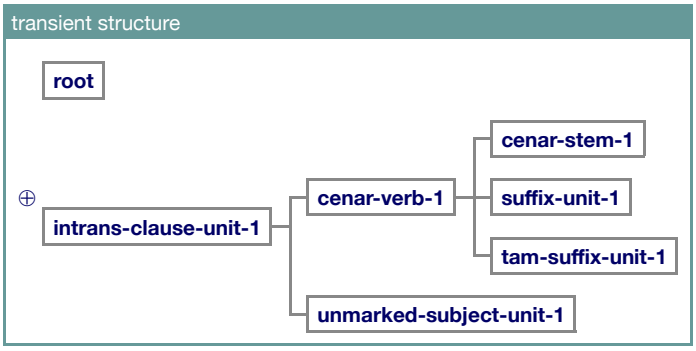
The grammar fragment that this demo describes targets the Spanish verb phrase, uniquely focussing on the conjugated verb form. Rather than storing all verb forms

individually as holistic chunks, the FCG grammar contains productive rules in the form of constructions, working together to build or analyse a single conjugation. For a simple form such as *cenabas*, 'you had dinner', the grammar makes use of six constructions to map the verb form to the following meaning representation:



Because we focus on the conjugation of individual verb forms, the grammar fragment only considers intransitive verb meanings. Therefore every verb, including transitive or ditransitive verbs, has only a single role in its meaning representation.

A constituency grammar approach is used to model conjugated verb forms in this grammar fragment. The resulting transient structure after producing the above meaning network looks as follows:



The utterance that can be extracted from this transient structure is *cen-aba-s*, where the three morphemes have been distributed over the three units under the `cenar-verb-1` unit. Yet, the transient structure contains more units than just these three form-bearing morpheme units. Three additional units have been created. The first one collects the stem morpheme and the two suffix morpheme units into a verb unit, containing agreement information and the tense-aspect-mood characterization of the form, together with the meaning of the form and the order of the morphemes (in the form feature). Another unit is made for the unmarked subject, which is a second singular person in this case. Together, the subject and the verb unit are the constituents of the intransitive clause unit.

A transient structure such as this one is the result of a repeated application process of constructions. In the case of *cenabas*, six constructions did their work and they can be split up into four different types: (i) Lexical constructions, (ii) stem constructions, (iii) suffix constructions and (iv) grammatical constructions.

3.1.1 Lexical constructions

The lexical construction set contains those constructions that capture the lemmas of verbs, similar to how they would be encountered in a dictionary, be it with annotations such as semantic, phonological and syntactic information. Figure 4 includes an example of a lexical construction for the verb *cenar* 'to dine, to have dinner'. A lexical construction for a verb lemma will always consist of two units on the conditional part: the `?cenar-verb` unit and the `?cenar-stem` unit. The stem unit is going to be a constituent of the verb unit, together with potential suffixes (to be added later). In production, all this lexical construction matches on is the meaning feature in the root unit (accessed by the HASH

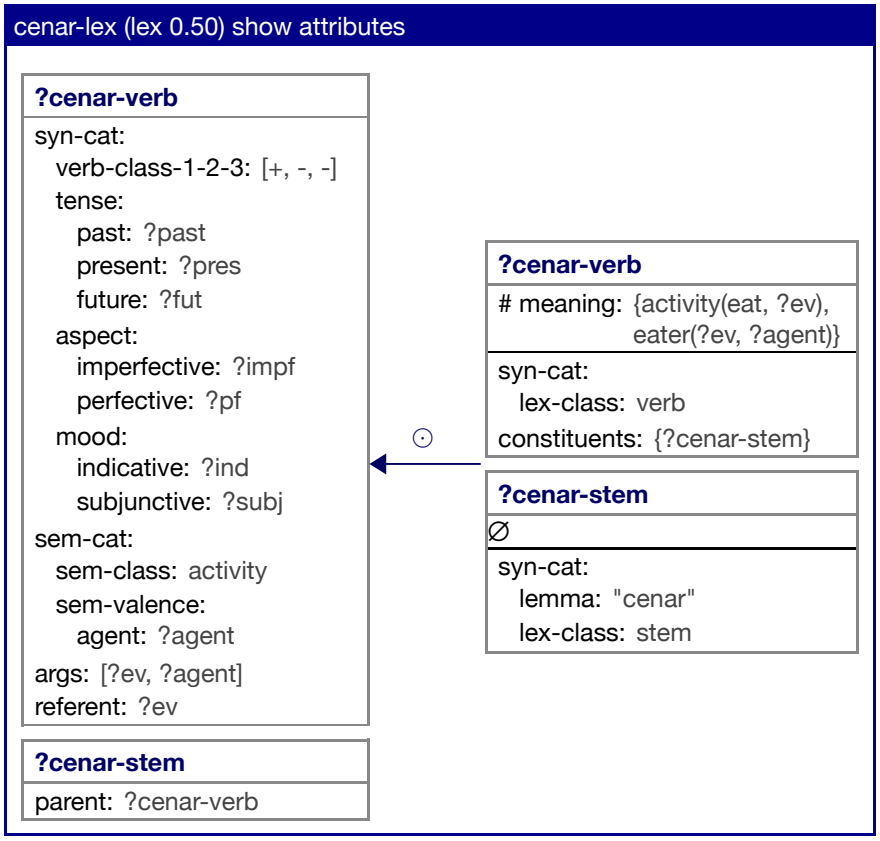


Figure 4: The lexical construction for the "cenar" lemma creates two new units in production: a verb unit with a stem unit as its constituent. It initializes a verb skeleton that will be filled in by other constructions, depending on which verb form needs to be realized. In comprehension, the verb and stem units are already present in the transient structure.

operator). In comprehension, on the other hand, we expect two units to be there already, linked by a constituency relation, with the stem unit bearing the lemma "cenar" and a `lex-class`. The contributing part will merge information into the two units, such as a verb-class (set to 1st verb class, since the lemma is ending on *-ar*), a sem-class, basic valency information and initialises the tense-aspect-mood features with variables. The semantic class is inherited from the meaning feature, and the agent of the activity is linked to the eater role.

3.1.2 Stem constructions

Lexical constructions do not contain actual word forms that will be encountered in sentences. This is the task of the stem constructions. For a default stem such as *cen-*, the corresponding stem construction maps the lemma "cenar" into a string feature "cen":

If you inspect the stem construction in Figure 5, you will see that there is a red feature inside the `phon-cat` feature. This colour highlighting indicates a negation of the `stem-realized` feature, meaning that the construction can only apply in production if the phon-cat of the stem unit does NOT contain a `(stem-realized +)` feature. Such a precondition is needed because this is the construction for the base stem of the verb, and thus constitutes a sort of default. Irregular stems are namely tried out first in production. If one of them would have been able to apply, the `cenar-base-stem` construction would no longer be triggered.

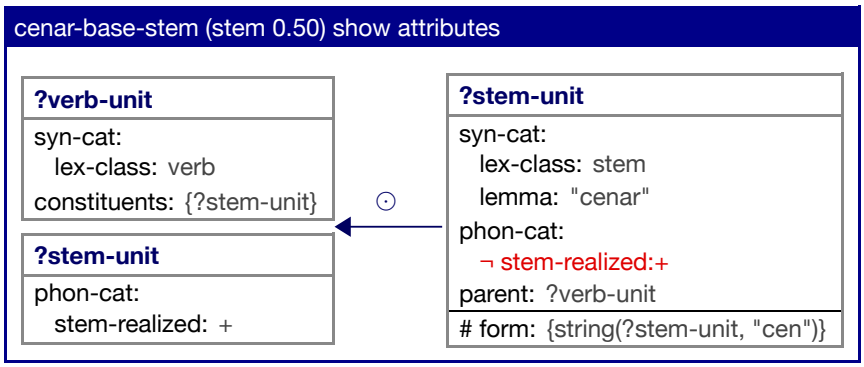


Figure 5: The default "cen-" stem, corresponding to the lemma "cenar", *to dine*.

3.1.3 Suffix constructions

To make the Spanish verb phrase modular, a verb form is analysed into three parts: a verb stem (see above), a tense-aspect-mood suffix and an agreement suffix (e.g.*cen -aba -s*). The *-abas* suffix is split into two parts: a morpheme that indicates the tense/aspect/mood of the verb form and a morpheme for the person/number information. The construction in Figure 6a indeed matches on a verb unit in the transient structure that has a past tense, imperfective aspect and indicative mood of the first verb class (in production); or an *-aba* suffix immediately adjacent to the verb stem (in comprehension). The contributing part of the construction creates a suffix unit, which is a constituent of the verb unit, and a sibling of the stem unit.

The person/number information is added by the `s-2sg-morph` construction (Figure 6b). Its preconditions in production are the following: a second singular agreement feature, a verb of any verb class that is not future and not perfective. It is also a requirement that the verb unit already has two constituent units. When these conditions are met, the construction makes the verb stem unit finite, as well as creating a second suffix unit that follows the verb stem unit (but does not have to be directly adjacent to it, hence the use of the `precedes` feature).

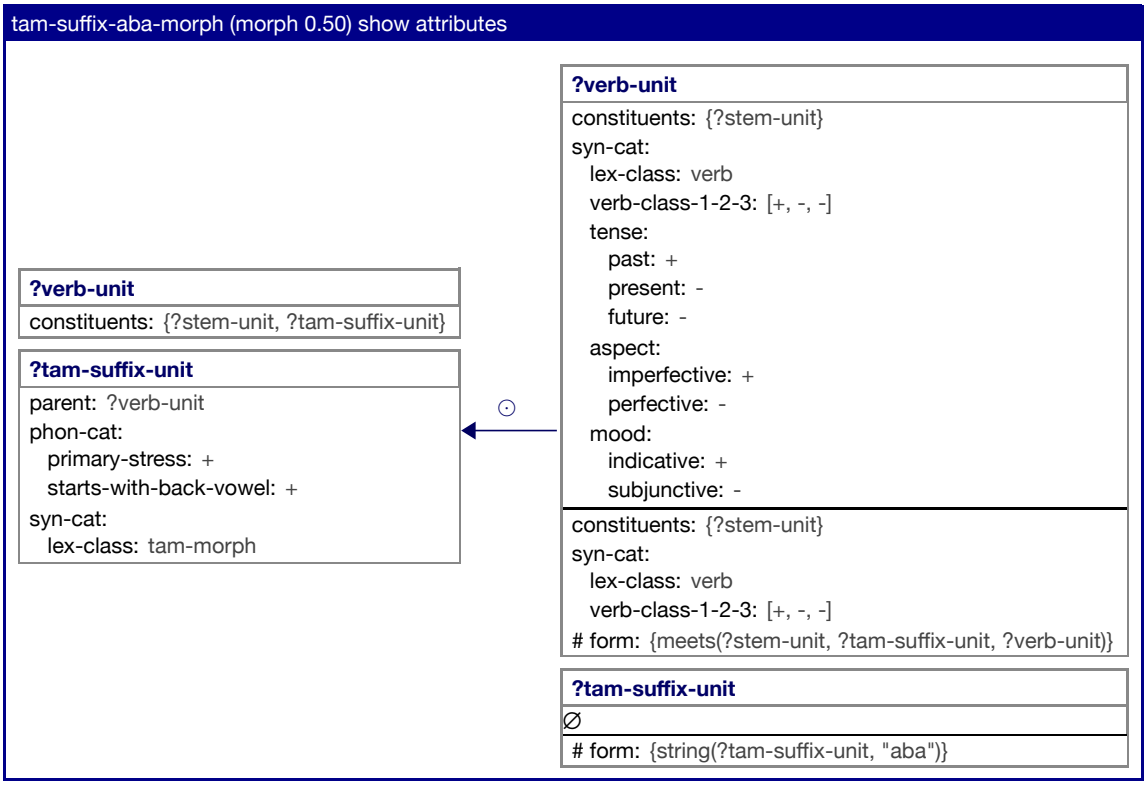
3.1.4 Grammatical constructions

Two remaining constructions are involved in the processing of a form such as *cenabas*: the intransitive second singular construction and the past imperfective construction. The first one matches on a verb unit with a second singular agreement feature in comprehension and merges a corresponding meaning representation into its structure (Figure 7). Also, a new unit, the VP unit is created that has two dependent units: the verb stem unit itself, as well as the unmarked subject unit.

Finally, the past imperfective constructions maps a meaning representation onto a syntactic configuration of the tense, aspect and mood features. No contributing part is required.

3.2 Verb conjugation

How do these constructions interact in an actual production setting? When producing the verb form *cenabas*, 'you had dinner', first in line are constructions that carve out a part of the meaning that needs to be expressed: the `cenar-lex`, the `2sg-covert-subject-cxn` and the `past-imperfective-indicative-cxn`. Then, the morphological constructions can elaborate this transient structure and attach actual forms to these meanings, through



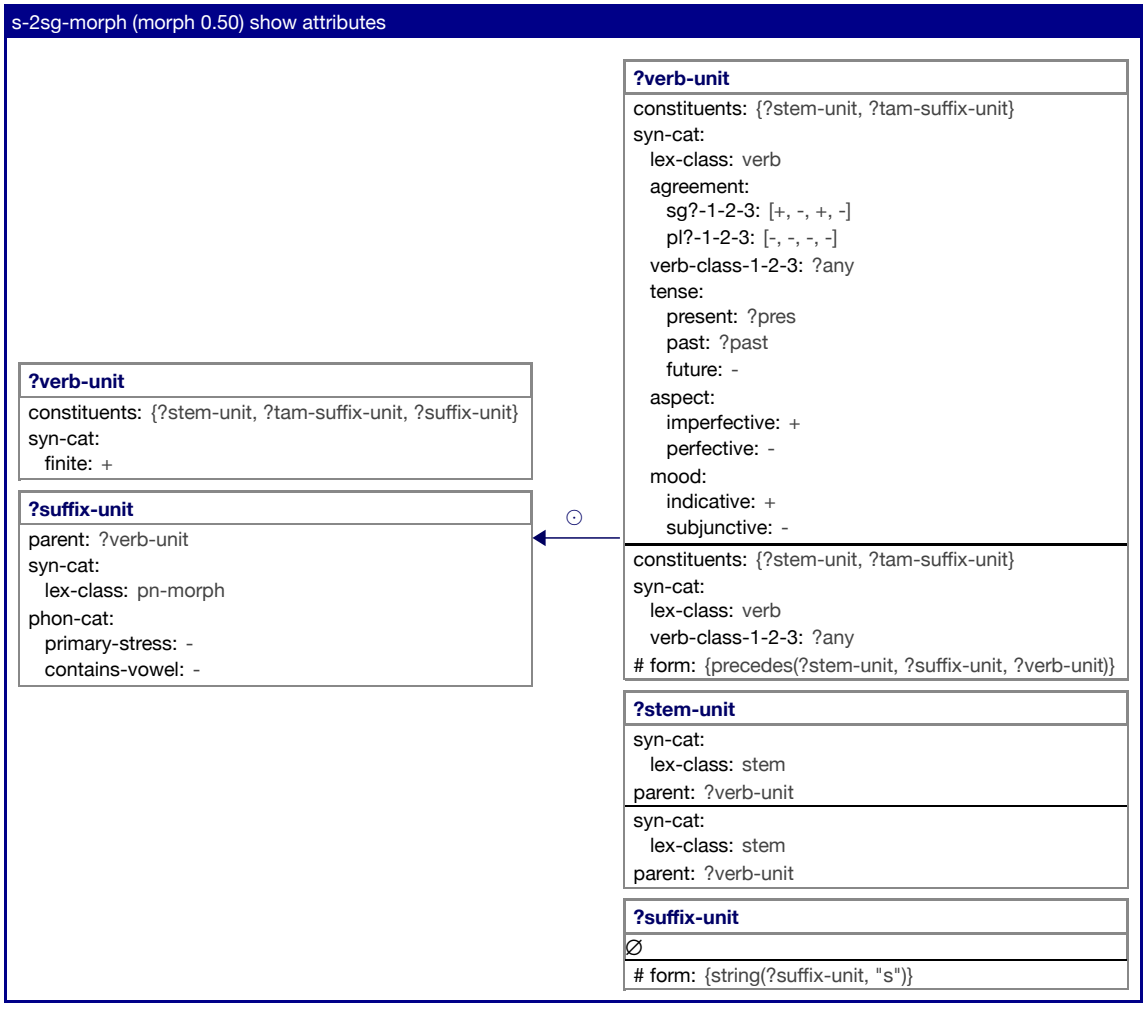


Figure 6: Suffix constructions add a new unit under the verb unit and attribute values to the syntactic features of tense, aspect and mood (a) or agreement (b).

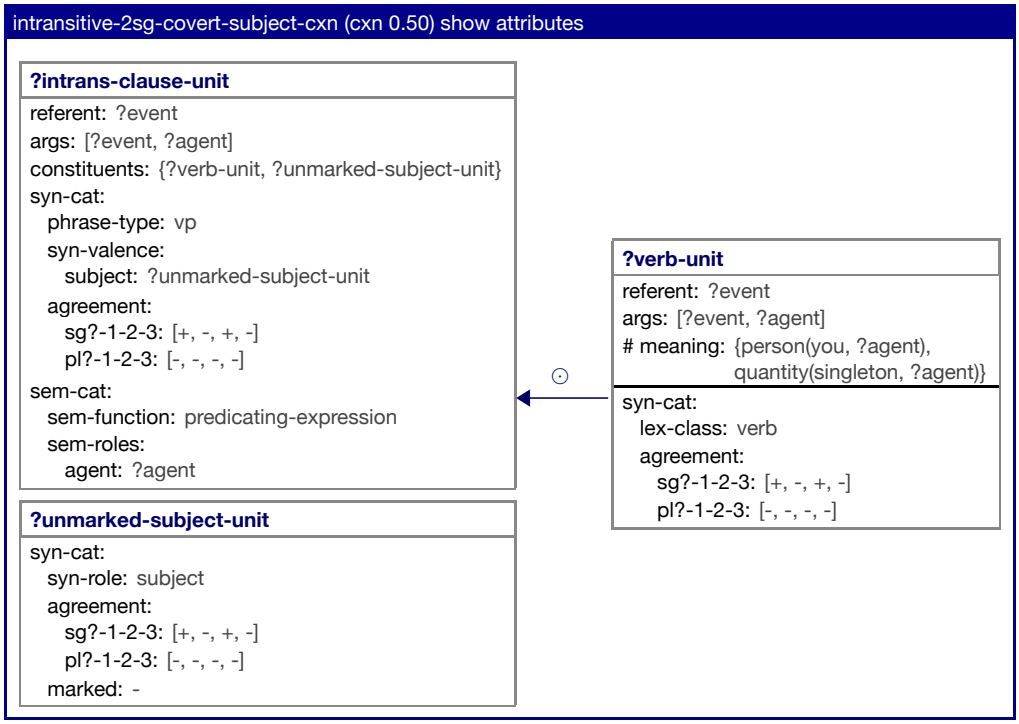


Figure 7: The intransitive construction for 2nd person singular agents.

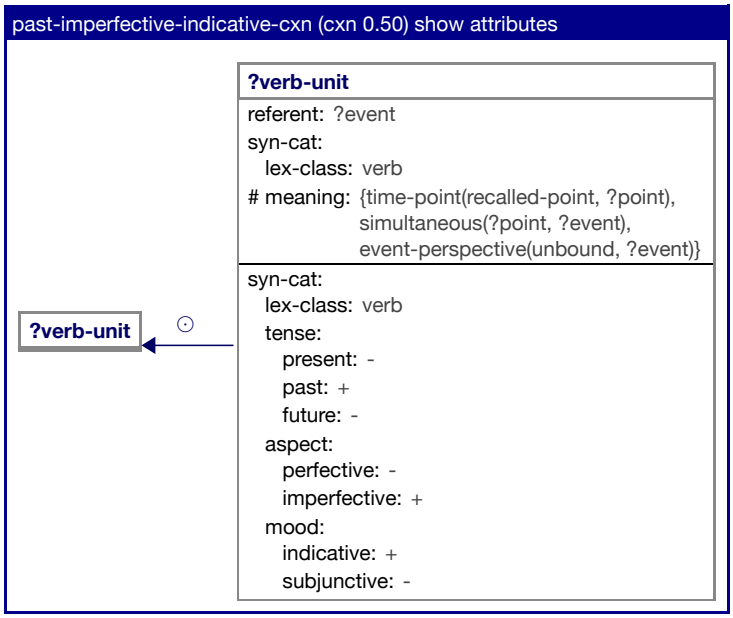


Figure 8: The past imperfective construction maps a certain conceptualization of time and event perspective into syntactic features for tense, aspect and mood.

the expression of certain syntactic features. The web demonstration that accompanies this paper demonstrates this application order and shows the effects of every construction on the transient structure that is being built.

Another feature of the grammar that is exemplified by the web demonstration is the comprehension process of syncretic verb forms such as *cen-aba*, which can both be used to say 'I had dinner' (1sg) or 'he/she had dinner' (3sg). In comprehension, the order of construction application is different from the one we observed in production. First the morphological constructions apply, as they match on string features that are present in the root unit. Then, the syntactic features that these constructions have added for tense and agreement for instance, are conditions for the grammatical constructions to apply. Two resulting meaning representations are found for the *cenaba* form. The lack of an overt person/number ending can both be interpreted as a first person singular or a third person singular. There is one construction that covers both cases: `no-marker-1/3sg-morph`. This construction is part of a special construction set (default), that is consulted only after the regular morphological constructions could not apply.

3.3 Stem changes

Verb conjugation in Spanish, as other Romance languages, is characterized by a large number of verb paradigms according to which specific infinitives are conjugated. Rather than simply combining the verb stem with one or two suffixes, the actual verb form can often be the result of transformations following from stress patterns or phonetic assimilation processes.

To account for multiple possible realizations of a verb stem (given the same lemma), a range of morphological allostructions needs to be defined, which compete with each other. For instance, for the verb *cocer* 'to cook', four stem allostructions are implemented, as visualized in Figure 9. Three irregular stems ("cuec-", "cuez", "coz-") compete with the base stem "coc-". Each of them have certain preconditions in production, such as a suffix unit that starts with a back vowel, and a stem unit that carries the primary word stress (cf. `cocer-cuez-stem`). Only when none of the irregular stems can apply, the regular `coc-base-stem` construction will trigger.

4 Towards a productive Spanish grammar

A basic grammar can and should be created by hand in Fluid Construction Grammar as the grammar engineer has to make decisions as to how certain constructions behave and work together in building or analysing prototypical conceptualisations or forms. Yet, to create large scale grammars that can be used in real-world applications such as language tutoring systems [Beuls, 2013] or dialogue systems, the original "seed" grammar needs to become productive to cope with new utterances that cannot yet be parsed or produced and internalise such novel uses so that they get incorporated into the construction inventory.

When a novel verb form is heard that cannot be parsed with the Spanish grammar at hand, its verb stem and its infinitive have to be retrieved in order to assign the form to a conjugation paradigm as they provide crucial information:

- A verb's stem (in combination with its verb class) gives away the conjugation type the verb belongs to (which will be encoded in the lexical construction's footprints feature).
- A verb's infinitive is an indicator of its verb class, a feature needed to select the appropriate tense-aspect-mood suffixes;

Yet, retrieving the infinitive from a conjugated verb form in Spanish is not always straightforward as many the paradigms of the second and the third verb class largely

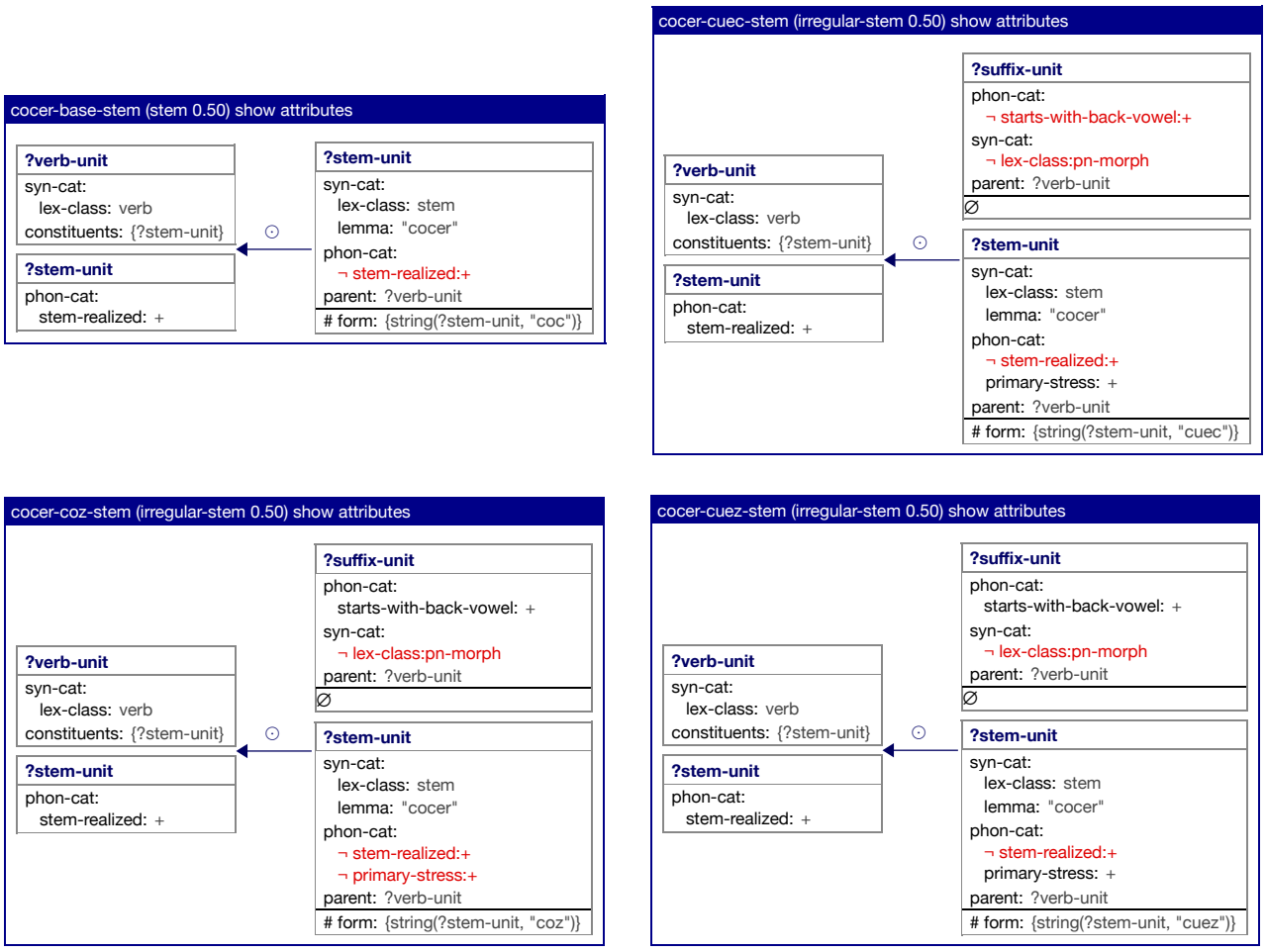


Figure 9: Four allostructions for the stem expressing the "cocer" lemma.

overlap in terms of the tense-aspect-mood and person-number suffixes they use. Moreover, a semantic annotation is needed to disambiguate between the indicative and the subjunctive mood as the opposite tense-aspect-mood suffixes are used: instead of *-a-* for the first verb class and *-e-* for the second or third (optionally *-i-*), the subjunctive mood requires the *-e-* suffix in the first verb class and the *-a-* suffix for the remaining classes (see also Figure 3).

A linguistically motivated decision tree designed by [Rello and Basterrechea, 2011] in the context of the development of the *Onoma* conjugator can decide on the particular conjugation paradigm an verb stem belongs to. As opposed to approximately one hundred and twenty conjugation models that are found in Spanish grammar books, they developed a set of nine patterns and a set of rules to decide on a verb's conjugation. Their description turned into a pedagogical tool for students of Spanish as a foreign language [Basterrechea and Rello, 2010]. The decision tree consists of just seven steps to identify the nature of any possible verb by reference only to its infinitive form[1].

On the one hand, productivity implies that newly learned verbs can be freely combined with entrenched constructions that look after their morphology in new situations. On the other hand, a truly productive grammar also allows for new phonological constructions to emerge. The next section explains how both types of productivity are included by making use of a manually coded decision tree inspired by the implementation of the Onoma Spanish conjugator. Finally, our open-ended grammar is tested on the full conjugational paradigms of the 600 most common verbs in Spanish.

4.1 Extending the seed grammar

Once the seed grammar has been carefully designed and implemented with the grammatical and morphological constructions needed to conjugate regular verbs, the next step is to couple a classifier to this grammar that is able to assign a particular verb conjugation template to a new infinitive. The same linguistically motivated decision tree (with minor modifications) as proposed by [Rello and Basterrechea, 2011] was implemented to find out the irregularity patterns of a particular infinitive. Figure 10 illustrates the main steps in the algorithm. Yet, many of the steps that make up the decision tree require external knowledge that needs to be provided, such as a list of auxiliary, copulative and primary verbs, a way to segment verb forms into prefix and stem as well as access to a dictionary to verify the chances that a verb changes its stem vowel through diphthongization or vowel raising (e.g. *contar* 'to explain' > *el cuento* 'the story';). The *could be irregular* outcomes in the decision tree need to be verified by means of such external knowledge.

Regular verbs go through the complete decision tree as they trigger a negative answer at every node. Certain verbs (including the already used example of *cocer* 'to cook') require a conflation of two verb paradigms. This is accounted for by the dotted arrows in the decision tree. Figure 10 hides the finer-grained distinctions needed to disambiguate between verb conjugation templates of the same irregularity type (i.e. that end up in the same 'yes' leaf). For instance, when answering 'yes' at the leftmost node *2nd/3rd conjugation* that follows *does its stem end on a vowel?*, there are five more checks needed to define the actual verb conjugation template:

1. Does its stem end on an "a"? → `caer`
2. Does its stem end on an "e" or "o" and is it a verb belonging to verb class 2? → `leer`
3. Does its stem end on an "e" and is it a verb belonging to verb class 3? → `reir`
4. Does its stem end on an "o" and is it a verb belonging to verb class 3? → `oir`

[1] Although, as noted by [Rello and Basterrechea, 2011, p. 2], "in some rare cases, external information which the system also provides is required".

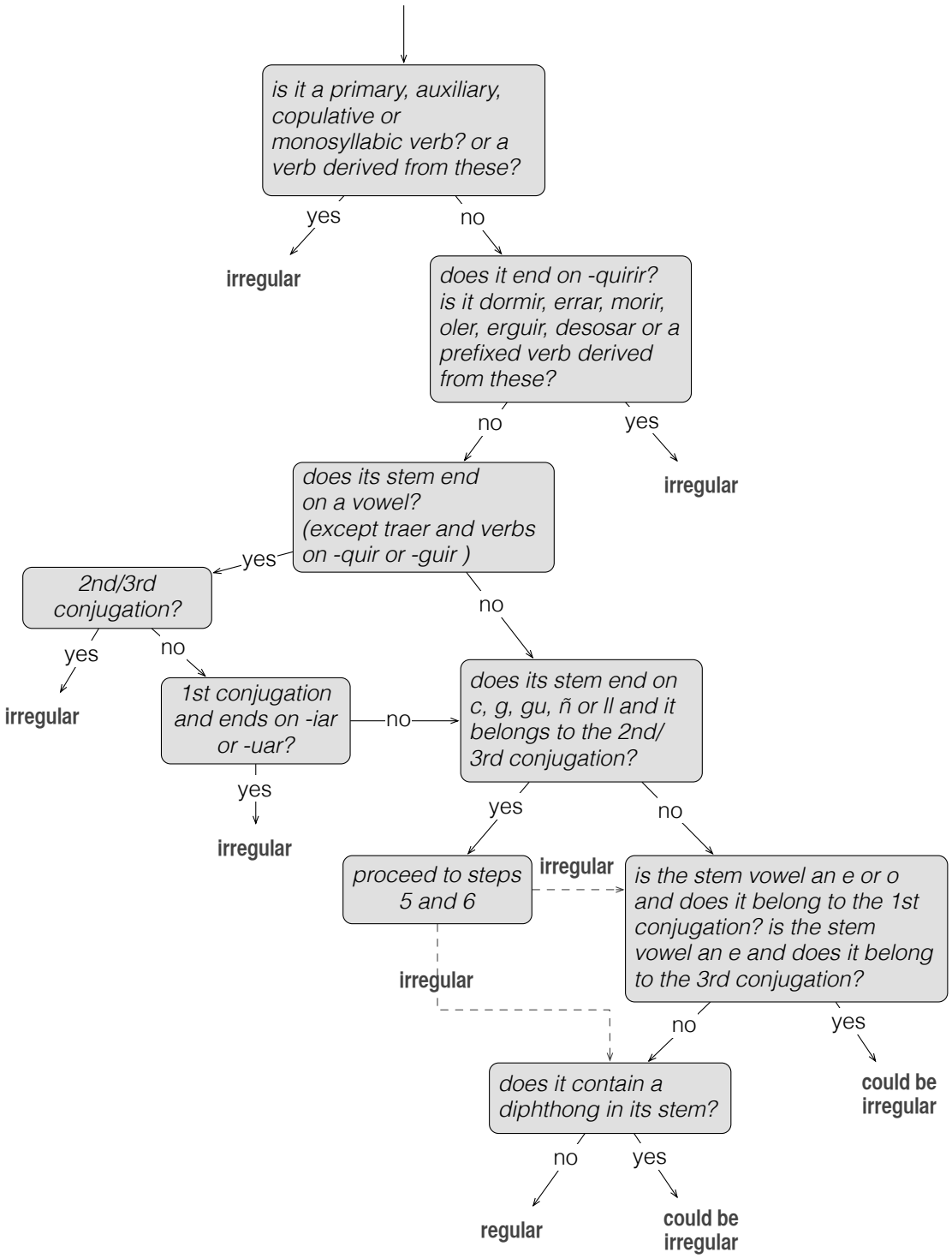


Figure 10: This decision tree decides whether a Spanish infinitive is regular or irregular on the basis of six main questions about its formal properties.

5. Does its stem end on an "u" and is it a verb belonging to verb class 3? → `huir`
   Does its stem end on an "ü" and is it a verb belonging to verb class 3? → `argüir`

The decision tree returns one or two conjugation types for a given infinitive. Based on this classification, new constructions can be added to the construction inventory. On the one hand, a lexical construction is added to capture the new infinitive, with its conjugation type(s) added to the `footprints` feature of the construction. Furthermore, a phonolog-

ical construction needs to be added that expresses the stem or suffix changes needed to transform the regular verb form to the appropriate form required by the specific conjugation type. For instance, when the verb *llegar* is learned by the construction inventory a lexical construction is added as well as the phonological construction for the `delegar` type, unless this construction already exists in the construction inventory. The knowledge required by these phonological constructions, that is the conditions under which certain changes take place, needs to be encoded in the grammar expansion algorithm.

4.2 Grammar evaluation

To create a large scale grammar on which the classification algorithm could be tested, I used a corpus that I received from Fred Jehle (University of New Mexico) with 11,467 conjugated verb forms, together with their infinitives. The conjugated forms in the corpus represent roughly the 600 most common verbs in Spanish. It was used on `http://users.ipfw.edu/jehle/verblist.htm`http://users.ipfw.edu/jehle/verblist.htm to present verb paradigms to students of Spanish, who could use it to look up an infinitive and retrieve its full conjugation.

Once all the infinitives from the Jehle corpus have been added to the FCG base grammar and the classification has been done, the resulting grammar has a size of 1575 constructions. Lexical constructions are by far the majority in this construction inventory with a share of 93% (1466 constructions), followed by phrasal constructions (21), morphological (35) and morpho-phonological constructions (53). There were thus no new phrasal or morphological constructions added in the grammar expansion phase. The classification tree only creates new lexical and morpho-phonological (phon) constructions. The high amount of lexical constructions is due to the number of irregular verbs that were conjugated, which resulted all together in 866 constructions, while only 564 infinitives were added to the grammar.

The majority of the 564 infinitives (58%) are regular verbs that were not classified according to a particular verb type and thus passed through the complete decision tree. The remaining 42% are irregular and semi-regular verbs that do not deviate from the regular verb conjugation paradigm on a number of verb forms. A total of 25 semi-regular verb types could be detected in the corpus data, with the most frequent verb type *secar* occurring 39 times. Seven of the verb types only had a single infinitive that was conjugated according to the type.

One way to verify whether the automatic verb classification is successful, is to parse a given verb form and reproduce it. When the result is the same as the original verb form, you know that the conjugation was correct. The evaluation tested 33954 verb forms, belonging to 566 verbs, and yielded on average 18 errors (5 runs). All these errors were due to ambiguous parses of verb forms that have a stem that can belong to any of two infinitives: sentir/sentar, regir/regar, presentir/presentar, sentir/sentar. Figure 11 contains an example of such an ambiguous parsing process where the verb stem *present-* can lead to *presentar* 'to present' or *presentir* 'to anticipate, to sense'. Because the testing function only takes one solution into account, one of the two lexical constructions that cover the stem *present-* triggers randomly. A solution to this ambiguity problem would be to implement a node test that checks whether the parsed meaning so far is still compatible with the situation that is being observed (the meaning). In a real language game, this checking is done automatically but in this verb form testing function, only the decontextualized verb form is available.

5 Conclusions and future outlook

This paper presented a way of implementing a compositional approach to Spanish verbal morphology in a computational Construction Grammar framework, namely Fluid Con-

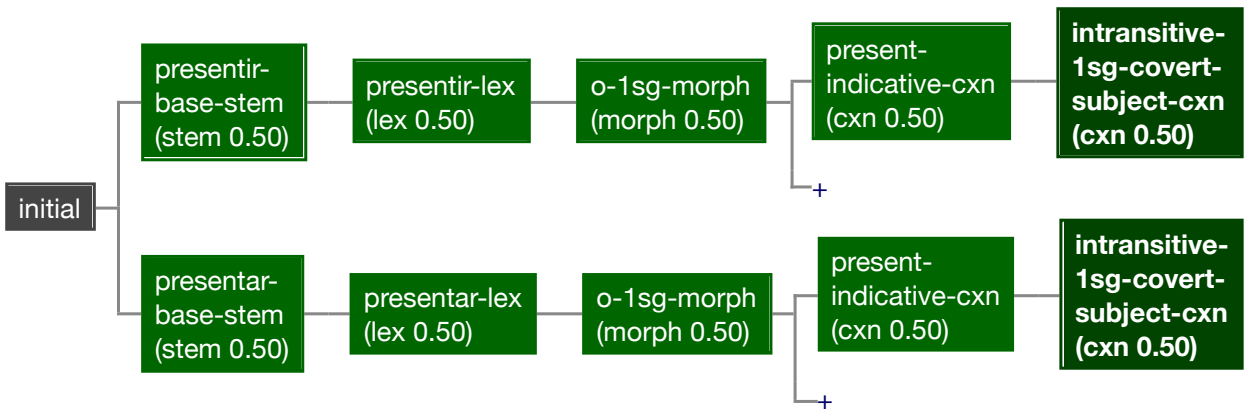


Figure 11: Parsing *presento* leads to two solutions that differ in the verb lemma that is used: *presentir* 'to anticipate' vs. *presentar* 'to present'.

struction Grammar. Rather than memorising complete verb forms in a memory-based approach [Daelemans and Van den Bosch, 2005], individual constructions handling verb stems, suffixes, phonological changes and time and aspect-related conceptualizations have been operationalized that capture the necessary generalizations to conjugate any Spanish verb. By first conjugating every verb according to its regular paradigm and then modifying the regular form accordingly when the verb belongs to a certain irregular or semi-regular conjugation type, the approach presented in this paper relates to the hypothetical form used by other conjugators that make use of finite state transducers in a two-level morphology approach [Koskenniemi, 1983]. In such an approach, the lexical level contains the stem plus affixes and the surface level the actual verb form (with potential stem changes). The application of constructions as shown is this paper is similar to a cascade of finite-state transducers but with a flexible representation that is coupled to a semantic space. Because a feature structure is much richer data-structure than a set of states with transition functions, our approach makes it easier to diagnose and repair learner errors because diagnostics can for instance signal when and why a suffix and a verb stem do not belong together (e.g. "cantías" instead of "cantabas" for *you sang* (impf.)) (see [Beuls, 2014] for a complete account of learner error correction with FCG).

To automatically increase the size of the grammar when a new infinitive is encountered, a decision tree was implemented that decided on the verb's conjugation paradigm, needed to learn a new lexical construction for the stem that contains a reference to the paradigm. Of course, a truly productive grammar that allows for language change (e.g. changes in stress patterns => ablaut verbs) should also be able to change the decision tree over time. We leave it to future work to explore this exciting route towards a fully creative grammar.

6 Acknowledgments

The results presented in this paper were originally part of my PhD dissertation. I therefore thank the jury members for their useful comments on explanations of the Spanish grammar, especially Alex Housen, Piet Desmet and Emmanuel Keuleers. A special word of gratitude goes to my supervisor Luc Steels for challenging me to always look for better solutions in grammar engineering. This research was funded by a basic research grant from the Flemish Agency for Innovation by Science and Technology.

References

[Basterrechea and Rello, 2010] Basterrechea, E. and Rello, L. (2010). El verbo en español. Molino de Ideas, Madrid.

[Beuls, 2011] Beuls, K. (2011). Construction sets and unmarked forms: A case study for Hungarian verbal agreement. In Steels, L., editor, *Design Patterns in Fluid Construction Grammar*, pages 237–264. John Benjamins, Amsterdam.

[Beuls, 2013] Beuls, K. (2013). *Towards an agent-based tutoring system for Spanish verb conjugation*. PhD thesis, Vrije Universiteit Brussel.

[Beuls, 2014] Beuls, K. (2014). Grammatical Error Diagnosis in Fluid Construction Grammar: A Case Study in L2 Spanish Verb Morphology. *Computer Assisted Language Learning*, 27(3):246–260.

[Bosque and Demonte, 1999] Bosque, I. and Demonte, V. (1999). *Gramática descriptiva de la lengua española* . Real Academia Española, Colección Nebrija y Bello, Espasa.

[Bull, 1965] Bull, W. (1965). *Spanish for teachers: applied linguistics*. R.E. Krieger Pub. Co., Malabar, Florida, USA.

[Comrie, 1976] Comrie, B. (1976). *Aspect*. Cambridge University Press, Cambridge.

[Croft and Cruse, 2004] Croft, W. and Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

[Daelemans and Van den Bosch, 2005] Daelemans, W. and Van den Bosch, A. (2005). *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

[Delbecque et al., 2001] Delbecque, N., Masschelein, D., and Vanden Bulcke, P. (2001). *El uso de los tiempos del pasado: Gramática española aplicada*. Wolters Plantyn, Mechelen.

[Gili y Gaya, 1943] Gili y Gaya, S. (1943). *Curso superior de sintaxis española*. Ediciones Minerva, Madrid, 1st edition.

[Koskenniemi, 1983] Koskenniemi, K. (1983). *Two-level morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki, Department of General Linguistics.

[Mateo, 1998] Mateo, F. (1998). *Les verbes espagnols*. Hatier, Montmorrilon.

[Mayol, 2003] Mayol, L. (2003). Acquisition of Irregular Patterns in Spanish Verbal Morphology. In *Proceedings of the Twelfth ESSLLI Student Session*, pages 1–11.

[Michaelis, 2006] Michaelis, L. A. (2006). Time and Tense. In Aarts, B. and McMahon, A., editors, *The Handbook of English Linguistics*, pages 1–24. Oxford.

[Ramsey, 1956] Ramsey, M. (1956). *A textbook of modern Spanish*. Holt, Rinehart and Winston, New York, revised by Robert Spaulding edition.

[Rello and Basterrechea, 2011] Rello, L. and Basterrechea, E. (2011). Onoma: A Linguistically Motivated Conjugation System for Spanish Verbs. Number 6608, pages 227–138, New York. CICLing 2011: The 12th International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science, Springer.

[Terrell and Hooper, 1974] Terrell, T. and Hooper, J. (1974). A semantically based analysis of mood in spanish. *Hispania*, 57:484–494.

[Whitley, 2002] Whitley, M. (2002). *Spanish/English Contrasts: A Course in Spanish Linguistics*. Georgetown University Press.