On auxiliary verb in Universal Dependencies: untangling the issue and proposing a systematized annotation strategy

Magali Sanches Duran
Núcleo Interinstitucional de Linguística
Computacional (NILC)
Instituto de Ciências Matemáticas e de
Computação, Universidade de São Paulo
(USP)
magali.duran@uol.com.br

Adriana Silvina Pagano
Faculdade de Letras, Universidade Federal
de Minas Gerais (UFMG),
apagano@ufmg.br

Amanda Pontes Rassi
Redação Nota 1000 Serviços educacionais
/ Somos Educação
amanda.rassi@somoseducacao.com.br,

Thiago Alexandre Salgueiro Pardo[1]
Núcleo Interinstitucional de Linguística
Computacional (NILC)
Instituto de Ciências Matemáticas e de
Computação, Universidade de São Paulo
(USP)
taspardo@icmc.usp.br

Abstract

Auxiliary verbs are universally recognized as components of verbal constructions. While there is no shortage of scholarship on these verbs in various linguistic traditions, uncertainty still remains on the best way to annotate them for Natural Language Processing (NLP) purposes. This paper reviews the evolution of the concept of auxiliary verbs to gather insights into forms of representing them in an annotation scheme and raises some issues with a view to leveraging the potential afforded by them in different NLP tasks. Using Brazilian Portuguese as an instance language and Universal Dependencies (UD) as annotation model, we argue for (i) annotating inflected verbs as heads, (ii) annotating auxiliary interdependence in an auxiliation chain; and (iii) adopting a more consistent treatment of auxiliaries to encompass tense, aspect, modality and voice in auxiliation chains. We further propose auxiliary type as a feature to be annotated which can be easily implemented in existing and new treebanks with substantial gains in enriching the information that can be extracted for different NLP applications.

1 Introduction

Thousands of years ago, writing ushered in a new era for mankind. The advent of writing made it possible for thoughts and information to be conveyed between individuals across distinct epochs and localities. The ideas recorded in writing began to fertilize other minds and generate new ideas, exponentially accelerating the evolution of ideas in human societies (Ridley, 2010).

As much as writing allowed knowledge transfer among individuals, it today supports the transfer of human knowledge to machines. This happens through Natural Language Processing (NLP), and one of the ways to train machines to process text and learn to extract from text much of what humans do with it is through annotated corpora. Corpus annotation has thus become a way to formally record the implicit and explicit linguistic knowledge that can be gathered from texts. The result enables NLP to develop statistical models for training new human language technologies (Ide, 2017).

Annotating a corpus is an undertaking that requires considerable effort in designing, executing, and reviewing an entire process. Since design involves the creation of models to represent linguistic information, the models are often reused in corpus annotation endeavors, both within a language and in languages other than the one for which the model was created.

One drawback of leveraging models is the fact that, because they represent facts of a language, they are to some degree dependent on the language the theory drew on for its study. The advantage of models' reuse, on the other hand, is that the use of the same annotation scheme becomes a means to compare languages. The comparison, in turn, makes it possible to create multilingual NLP applications. This is the aim of the Universal Dependencies (UD) model (Nivre 2015, Nivre 2020), which is designed to be language independent. At the time of writing, there are over 200 corpora annotated with the UD model in just over 120 languages.

The fact that many linguists and computer scientists use UD and strive to instantiate it in their languages has promoted numerous discussions around its guidelines. One such discussion revolves around auxiliary verbs. Some languages, for instance, have opted for tagging tense and passive voice auxiliaries as AUX; others include modal verbs under this tag, and still others add aspectual verbs to the set. While UD does not require languages to follow a single standard, it recommends that only strongly grammaticalized auxiliary verbs be annotated as such. This, in turn, raises further discussion as to where to draw the line for an auxiliary to be considered fully grammaticalized.

This paper grew out of a concern on how to best represent auxiliary verbs in UD scheme towards building a proposal to leverage the full potential afforded by them in different NLP tasks. Drawing on Brazilian Portuguese, we contend that there are substantial gains to be obtained from (i) annotating auxiliaries as heads; (ii) annotating auxiliary interdependence in an auxiliation chain; and (iii) adopting a more consistent treatment of auxiliaries to encompass tense, aspect, modality and voice in auxiliation chains. We further propose auxiliary type as a feature to be annotated for the purpose of enriching information to be tapped from treebanks. Our proposal offers several benefits to NLP tasks, such as enhancing detection of subjects for information extraction by annotating inflected verbs as heads; temporal reference detection by relying on tense and aspect auxiliaries; and speculation detection by leveraging modal auxiliaries as cues for that task.

In Section 2, we briefly review the evolution of the concept of auxiliary verbs, highlighting the points that are important to our discussion. Section 3 discusses the auxiliation process and provides examples in Brazilian Portuguese for four types of auxiliaries: tense, aspect, modality and passive voice diathesis. In Section 4, we exemplify ways to annotate auxiliary verbs in UD, discussing their pros and cons and presenting a proposal to reduce different forms of annotation to an interpretation of auxiliaries common to all of them. Section 5 concludes our study.

2 Contributions over time towards the concept of auxiliary verbs

The topic of auxiliary verbs has been extensively discussed in the last decades, a thorough review being outside the scope of this paper. We will focus hence on the works that most contributed to advancing discussions of the auxiliary verb concept, since this concept is fundamental to the decisions about the UD annotation scheme we would like to argue for.

Our review begins with Tesnière (1959), an author who compared auxiliaries to free morphemes, though ascribed to them a distinctive nature of being inflectional. Auxiliary verbs, for Tesnière, help other verbs enable a subcategory transfer (in his account, of tense and voice) and are totally devoid of semantic content. Auxiliaries are classified as compound verb forms, as opposed to simple forms, and operate with auxiliated ones. Tesnière described auxiliary verbs as those assuming grammatical functions whereas auxiliated verbs contribute with the semantics. Being acquainted with English grammar (he mentions the verb *do* as an auxiliary), he admitted other functions of auxiliaries, which he suggested when he used *etc*. in: "One distinguishes between auxiliaries of tense (past, future), auxiliaries of voice (passive), etc." (Tesnière, 1959, p. 403).

Benveniste (1974) elaborated on Tesniere's description, recognizing verbal chains of auxiliaries of tense, modality and diathesis (voice). The author coined the term *auxiliation* to refer to a process that syntagmatically joins an auxiliating form to an auxiliated one, avoiding the use of the term *auxiliary*. For simplicity, we adopt the term *auxiliary verb*, even when referring to Benveniste's work.

By including modal verbs as a further type of auxiliation, Benveniste evidenced that auxiliation chains are longer and more complex than previously believed; yet he did not include aspectual verbs among auxiliaries. Another important contribution by Benveniste was to show that there is sequential order for auxiliary verbs to occur, namely, modal - temporal - passive voice - full verb, within a process he called *over-auxiliation*. Despite acknowledging that an auxiliary verb is the verb that takes person, number, mood, and tense inflections in a compound form and showing compound forms made up by up to three types of auxiliary functions (tense, modality, and passive voice diathesis), Benveniste did not expand on the fact that, in the case of longer chains, auxiliaries after the first one do not take inflections. Neither did he explicitly state that the second auxiliary in a chain is auxiliated by the first one and so on. He did state, however, that an auxiliary of passive voice diathesis is always the last one in a chain before a full verb, because no auxiliary verb can undergo passive voice diathesis.

Another relevant contribution to the concept of auxiliary verbs was made by studies of grammaticalization, mainly after the 1990s. Heine (1993) does a survey of the different ways languages express features of tense, modality and verbal aspect, pointing out that the lack of agreement around a concept of auxiliaries is largely due to the diversity of phenomena. For Heine, one of the sources for linguists' disagreement can be traced to Chomsky's AUX, a universal category he introduced in 1956, which is in fact not directly related to auxiliary verbs. Heine's review shows that auxiliary verbs have been at times considered as main verbs, as non-autonomous verbs, and still as a different grammatical category of verb altogether. Likewise, in dependency grammars, auxiliary verbs are usually considered as dependent by some authors while others posit them as heads. Steele (1994, p.818) praised Heine's work for his survey of views on auxiliary verbs, but criticized him for not tackling issues such as which verb is head and which one is dependent in dependency relations.

Kuteva (2001), who set out to complete the work of Heine, remarked that the big problem is the fact that some linguistic traditions disregard the dynamic character of the process of auxiliation, which prevents new auxiliaries arising in languages from being recognized. For her, auxiliation is an ongoing process and auxiliary verbs can be found at various stages in this process. There is thus no limited set of auxiliary verbs and one cannot separate auxiliary verbs from the verbs that gave rise to them.

Andersen (2006) agrees that auxiliation is a dynamic process, “so the class is continually losing and acquiring new members” (p.4). The author compares auxiliary constructions in over 800 languages and concludes: “There is no, and probably cannot be, any specific, language independent formal criteria that can be used to determine the characterization of any given element as a lexical verb or an auxiliary verb.” (p.5). He makes an important distinction between inflectional and semantic heads. The former encodes features responsible for making the construction to be grammatical, whereas the latter determines valence (argument structure). In some languages, the inflectional and the semantic heads are conflated, as the full verb is the one bearing inflections. In others, the inflectional head is the auxiliary and the semantic head is the full verb. Therefore, depending on the annotation purpose, dependence relations may prioritize the inflectional or the semantic head[1].

Krug (2012) discusses the grammaticalization of auxiliary verbs and illustrates the process of full verbs becoming auxiliaries for the English language. According to the author, and for this he draws on Bolinger (1980, apud Krug, 2012), it suffices for a verb to receive a complement in infinitive form to enter a path of grammaticalization. Krug cites the following characteristics of auxiliaries:

- they may coexist with a full homonymous verb;
- they contribute to expressing tense, aspect and modality (known as TAM);
- they do not occur alone, except in cases of elliptical full verbs (easily recoverable in context);
- they are complemented by verbs in non-finite forms (gerund, participle and infinitive).

[1] In fact, Andersen (2006) basically recognizes three patterns of auxiliary verb construction inflections: AUX-headed (the auxiliary is the inflected verb), which is the most common pattern; LEX-headed (the full verb is the inflected verb, as in Eneats, Bulgarian, Macedonian, Hatam, Koiari and Kwerba); and doubled inflections (both auxiliary and full verbs inflect, as in Gutob, Mombelo and Mumbami).

Krug acknowledges the fact that auxiliaries for passive voice diathesis and negative and interrogative constructions are considered by some linguists, but he does not include them in his account.

3 Discussion

Prior to Benveniste, an auxiliary construction was seen as a set of verbs complementing each other: one of them expressing grammatical features and the other semantic ones. Benveniste posited *over-auxiliation* and included modal verbs in the auxiliation process. The chain of over-auxiliation, including aspectual verbs, conforms to the following sequence of occurrence: modal - temporal - aspectual - passive voice - full verb (see Figure 1). Over-auxiliation is of key importance for corpus annotation, though it is still under-explored in accounts on the matter. Bearing in mind that each auxiliary verb imposes a non-finite form on the auxiliated verb, we can argue that, except for the first auxiliary, which holds inflections, all the other auxiliaries in a verbal chain are, concomitantly, auxiliated by a preceding verb and auxiliary to the following one, as shown in Figure 1. This implicates that the traditional labor division ascribed to auxiliary verbs as representing grammatical functions and full verbs as representing semantic functions cannot be sustained.

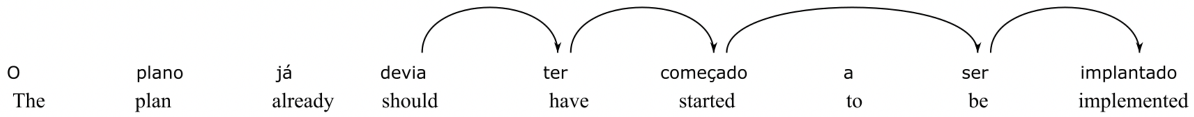

Figure 1: Auxiliary chain showing over-auxiliation process

Verbs in Figure 1 can be thus analysed:

- **devia [dever (should)]**, auxiliary of modality, requires the auxiliated to be an infinitive form; therefore, the verb auxiliated by *devia* is the verb *ter*.
- **ter [ter (have)]**, auxiliary of tense, requires the auxiliated to be a past participle form; the verb auxiliated by *ter* is the verb *começar*. *ter* is auxiliated by *devia* and is auxiliary to *começar*.
- **começado [começar (start)]**, auxiliary of aspect, requires the auxiliated to be an infinitive form and to be introduced by the preposition *a*; the verb auxiliated by *começar* is the verb *ser*. Therefore, *começar* is auxiliated by *ter* and is auxiliary to *ser*.
- **ser [ser (be)]**, auxiliary of passive voice, requires the auxiliated to be a past participle form. The verb auxiliated by *ser* is the verb *implantar* (past participle: *implantado*), which is the full verb in this sentence. Therefore, *ser* is auxiliated by *começar* and auxiliary to *implantar*.
- **implantado [implantar (implement)]** is a full verb, auxiliated by *ser*.

A productive way to explore the concept of auxiliation is to focus on the concept of auxiliated verb rather than on the concept of auxiliary verb. An auxiliated verb may be an auxiliary or a full verb. An auxiliated verb is a verb that takes the non-finite form required by its auxiliary, is introduced by the preposition (if any) required by its auxiliary, and has the same subject as its auxiliary. Therefore, we may have an auxiliation chain whenever all verbs in a chain share the same subject. However, we cannot affirm that the verbs in a chain sharing the same subject are auxiliaries followed by a full verb, as it depends on which verbs will be considered auxiliary in each work and for what purpose. In auxiliation verbal chains, the first verb is only auxiliary and the last verb is only auxiliated (full verb), but the verbs in between are both auxiliary and auxiliated, which shows that these two categories are not mutually exclusive.

The question to be posed is not whether a verb is an auxiliary, but whether it is an auxiliary in a given chain and what can be leveraged from its annotation. This approach makes it possible to overcome the much debated need to define a list of auxiliary verbs. Discussions about whether a verb is an auxiliary or not have always been based on comparing verbs with prototypical auxiliaries, i.e., those whose grammaticalization process is well advanced. In English, some auxiliaries (*can*, *may*, *might*, *should*) are fully grammaticalized to the extent that they do not compete with homonymous full verbs, do not require *to* to introduce the auxiliated verb, and do not require another verb to construct a negative and an interrogative form[2]. In other languages, such as Brazilian Portuguese, auxiliary verbs are at different stages in the grammaticalization process.

Several scholars have tackled the task of describing auxiliary verbs in Brazilian Portuguese. Among them, Pontes (1973) and Lobato (1975) are two particularly exhaustive accounts, each exploring likely criteria to classify a verb as an auxiliary one. Pontes (1973) takes a syntagmatic view on auxiliaries and discusses interdependence relations between auxiliary and auxiliated verbs. Lobato (1975) performs different probes to try to differentiate auxiliary from non-auxiliary verbs. More recently, Ilari & Basso (2014) systematize criteria to assign auxiliary status to a given verb, considering all instances found in corpora regardless of the degree of grammaticalization a verb is still exhibiting. When it comes to grammar textbooks, lists of auxiliary verbs can be found in most of them, no two lists being alike, which shows the variety of criteria and stances taken by grammarians in Brazil.

Drawing on Portuguese as a sample case, we would like to argue for a view on auxiliary verbs as verbs in their own right, implicating that a verb can be an auxiliary verb in some uses and a full verb in others, in the latter operating to help construe a variety of meanings. While they may be semantically weak, auxiliary verbs have a strong role in the syntax of the clause, its finite form agreeing with the subject and dictating the form of their auxiliated verbs. Auxiliary verbs can take part in a chain of several auxiliary verbs and be modified by adverbs, which sets them apart from fully-grammatical words. To better grasp the behavior of each auxiliary type in Brazilian Portuguese, we will briefly address four main groups (tense, modality, aspect and passive voice diathesis) and their characteristics below.

3.1 Auxiliaries of tense

Brazilian Portuguese expresses tense basically through morphological desinence. The so-called tense auxiliaries *ter* and *haver* are used to express a previous past event within the past itself (1) and a previous future event within the future (2). For this reason, *ter* and *haver* are tense auxiliaries in some environments only, as in the following examples.

(1) Quando olhei, ele já **havia atirado**. (When I looked, he **had** already **shot**.)
(2) No dia que você vier eu já **terei partido**. (The day you will arrive I **will be gone**.)

As seen above, both *ter* and *haver* require the auxiliated verb to be a past participle form. However, a past participle form is not a criterion sufficiently strong to single out occurrences of *ter* and *haver* as tense auxiliaries. *Ter* may combine with a past participle form in other tenses to express aspect (3) and resultative constructions[3] (4).

(3) Ele **tem vindo** aqui todos os dias. (He **has been coming** here every day.)
(4) Ele **teve aprovado** seu visto só ontem. (He **had** his visa **approved** only yesterday.)

In the case of *haver*, this verb may be followed by a past participle form, which is not actually a verb, but a noun. In such cases, *haver* is not an auxiliary verb. This is the case of (5) and (6):

(5) Não **houve** **comunicado** prévio dos organizadores
Not had communicated[4] prior of the organizers
There was no prior notification by the organizers.

[2] Among less grammaticalized auxiliaries in English, Osborne & Gerdes (2019) point out *be going to.*

[3] Resultative constructions resemble a kind of diathesis where the subject has the semantic role of benefactive. Diathesis is the alternation of semantic roles: in the passive voice diathesis, the patient is the subject; in the causative diathesis, the cause is the subject.

[4] The word *comunicado* is the past participle of the verb *to communicate* and means both *communicated* and *notification*. Many past participles in Portuguese are employed as true nominals.

(6)	Não	**há**	**sentido**	em	fazer	isso
	Not	have	felt[5]	in	to do	this

There is no sense in doing that.

In examples 5 and 6, *comunicado* and *sentido* are nouns and not verbs. In both, *haver* construes existence and is inflected in present and perfect tenses, i.e., it is not an auxiliary of tense.

3.2 Auxiliaries of modality

Brazilian Portuguese has two modal verbs that are more highly grammaticalized than others: *poder* and *dever*. Both express several types of modality: permission, obligation and possibility. *Poder* has no homonymous full verb, but *dever* does [*dever* (*to owe*)]. There are several less grammaticalized modals like, e.g., *tentar* (*to try*). Some of them can be used as full verbs as is the case of *saber* (*to know*) and some of them, as is the case of *pretender* (*to intend*) and *querer* (*to want* or *would like to*), can also take a finite clause as a complement, its subject not being the same as the one of the main clause. However, whenever followed by an infinitive, those verbs share the same subject. These two possibilities are illustrated by examples (7) and (8).

(7) Você **quer marcar** uma consulta semana que vem? (**Would you like** to **schedule** an appointment next week?)

(8) Você **quer** que eu **marque** uma consulta semana que vem? (**Would you like** me to **schedule** an appointment next week?)

The particular behaviour of modal verbs mentioned above tends to exclude such verbs from traditional lists of modal verbs in Portuguese[6]. However, for NLP, verbs like *pretender* (to intend), *querer* (to want), *saber* (to know), *tentar* (to try), etc. followed by an infinitive verb are important cues for deducing whether an event has occurred or whether a statement is a fact or mere speculation.

Modal verbs in Brazilian Portuguese are in a stage of grammaticalization in which they have not lost their semantic load, since even the most grammaticalized one, *poder*, is polysemous: it construes permission (9) or probability (10). The same holds for *dever*, which construes obligation (11) or probability (12):

(9) Você **pode entrar**, se quiser. (You **may come** in if you want.)

(10) **Pode chover** hoje à noite. (It **may rain** tonight.)

(11) O funcionário **deve usar** uniforme todos os dias no trabalho. (Employees **must wear** their uniform every day at work.)

(12) O atraso **deve ser** por causa da chuva. (The delay **must be** due to the rain.)

Grammaticalization studies point out that there may be verbs at various stages in the grammaticalization process regarding their use as auxiliaries and this seems to be the case for many modal verbs in Brazilian Portuguese.

3.3 Auxiliaries of Aspect

Aspectual verbs express how events occur in time. The meaning of aspect can be readily grasped through examples of some of its subcategories: frequentative (informing an event repeats frequently), inchoative (informing an event has started) and terminative (informing an event has finished).

[5] The word *sentido* is the past participle of the verb sentir (*to feel* and means both *felt* and *sense*.

[6] In fact, some modal verbs such as *querer*, *desejar*, *pretender* are classified as full verbs realizing mental processes in systemic-functional descriptions of Portuguese. Likewise, in some accounts on auxiliary verbs, modals do not fulfill all the criteria to be considered auxiliary verbs (cf. LOBATO, 1975).

There are aspectual verbs in Brazilian Portuguese that convey information on the event. For example, the aspectual verb *chegar a* (literally *arrive to*) followed by an infinitive signals the event took place some time ago and lasted for a while, but did not persist.

(13)	Ele	**chegou**	**a**	**pensar**	em	abandonar	o	Brasil
	He	arrived	to	think	in	to abandon	the	Brazil

He **even thought** about leaving Brazil for good.

If we did not take into account the dynamic character of the process of auxiliation, as pointed out by Kuteva (2001), we would not be able to recognize new aspectual verbs arising in Portuguese. For example, the verb *dar de* [*dar* (give)], followed by infinitive, informs the event has become a habit:

(14)	Ele	**deu**	**de**	**assistir**	filmes	de	terror	ultimamente
	He	gave	of	watch	movies	of	horror	lately

He **took to watching** horror movies lately.

(15)	Eu	**dei**	**de**	**suspeitar**	de	todo	mundo	depois	que	fui	enganada
	I	gave	of	suspect	of	all	world	after	that	was	deceived

I **became suspicious** of everyone after I was deceived.

Besides being prolific, aspectual verbs are the least grammaticalized verbs in Brazilian Portuguese. Some compete in interpretation with full verbs, as is the case of *acabar de* [*acabar* (finish)], which, followed by an infinitive introduced by the preposition *de*[7], is aspectual in (16) and full verb in (17).

(16)	O	filme	**acabou**	**de**	**começar**
	The	movie	finished	of	to start

The movie **has just started**.

(17) Ele já **acabou de ler** o livro. (He has already finished reading the book.)

3.4 Auxiliary of passive voice

In Portuguese, the passive voice may be constructed by the auxiliary verb *ser* (to be) followed by a past participle (which we call analytic passive voice) or by adding the pronoun *se* to a transitive verb (which we call synthetic passive voice). In analytic passive voice, the subject is the prototypical patient and comes to the right of the verb *ser*:

(18) As cartas de sentença **foram assinadas** pelo juiz. (The sentencing letters **were signed** by the judge.)

Unlike the auxiliary verbs of tense, which also require the participle form of their auxiliated, the participle of the passive voice is not invariable: it agrees in number and gender with the subject of the passive voice (in Portuguese, number and gender are typical inflections of nouns, while the typical inflections of verbs are mood, person, number, and tense). This fact enables us to verify agreement between verb and subject regardless of which verb is the head of the subject dependency relation.

In Portuguese, only the passive voice auxiliary is fully grammaticalized and may occur in all verb tenses. Tense auxiliaries are well grammaticalized, but need to be annotated as such, though only in some verb tenses. Modals and aspectual verbs are less grammaticalized, but this does not mean their annotation is less important for NLP.

[7] One feature of aspectual verbs that poses a challenge to their annotation as auxiliaries in UD is the fact that most of them require a preposition to introduce the auxiliated verb. UD does not provide a specific dependence relation to link this preposition to one of the verbs (since the preposition neither marks case, nor introduces a subordinate clause). We have opted for annotating prepositions for verb and noun arguments as ADP.

4 Auxiliaries in Natural Language Processing

The different views among linguists on which verb categories can actually be considered auxiliary pose an additional challenge to model them for the purpose of annotation in NLP.

For syntactic annotation based on constituencies, the key issue is to decide which of the verbs in a verbal phrase is the head. For annotation based on dependencies, the decision encompasses which verb is head and which one is dependent, as well as which dependency relation links each of the verbs to the others.

In UD, there is a Part of Speech (Pos) tag for full verbs (VERB) and a PoS tag for auxiliary verbs (AUX). Its guidelines leave it up to each language to define which categories of auxiliaries will be annotated as AUX and which verbs are prototypically used in each category. UD basically recommends that the auxiliary verbs specified in the annotation guidelines should be highly grammaticalized in the language. This has encouraged very conservative decisions, so that not all languages annotate modality auxiliaries (modal verbs) and aspect auxiliaries (aspectual verbs), as they tend to be less grammaticalized than other auxiliaries.

Automatic identification of auxiliaries in Portuguese has already been focused on by Baptista et al. (2010), who consider an extensive list of over 26 auxiliary verbs; however, despite aiming at a dependency parser, the authors do not follow UD guidelines.

In Portuguese, one of the main probes for identifying the subject is through verb agreement. The verb holding verb inflections is thus a natural candidate to be the head of the **nsubj** relation. This has implications in cases where there are auxiliaries together with a full verb. In UD, once a verb is annotated as AUX, there is no possibility to annotate it as a head: it has to be dependent in an **aux** dependency relation. Only verbs annotated as VERB may be heads of dependency relations.

Therefore, when an auxiliary verb, annotated as AUX, happens to be the first in an auxilation chain, thus keeping the inflections, a direct dependency relation between the inflected verb and the subject is not annotated, which precludes easy extraction of subjects. Still, if a verbal auxiliation chain contains several verbs annotated as AUX, the distance between the subject and the head (full verb) is longer and, as the full verb is always in an infinite verb form, the main cue to verify agreement between subject and verb can be missed. Most importantly, once a verb annotated as AUX cannot be head in a dependency relation, this prevents annotating over-auxiliation, which is marked by the non-finite verb form required from an auxiliated by its auxiliary. This is a major drawback, implicating that rich morphological and syntactical information is left untapped for linguistic studies and NLP applications.

As far as we can see, there are three options to annotate auxiliation chains using dependency relations, which are illustrated in the three figures below.

Figure 2 shows annotation of all verbs with the PoS tag VERB, the first one being the head and the second one a dependent in a **xcomp** dependency relation, this annotation being iterated between the following verbs in the chain.

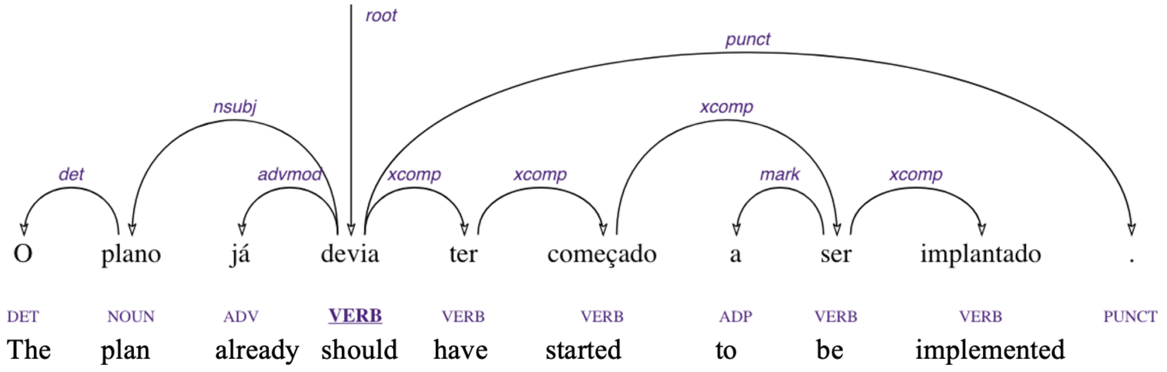

Figure 2: Annotation of inflected verb (auxiliary) as head

Figure 3 shows annotation of the more grammaticalized verbs as AUX (tense and passive voice auxiliaries) and the less grammaticalized ones as VERB (modal and aspectual auxiliaries), obtaining a combination of dependency relations **aux** and **xcomp**.

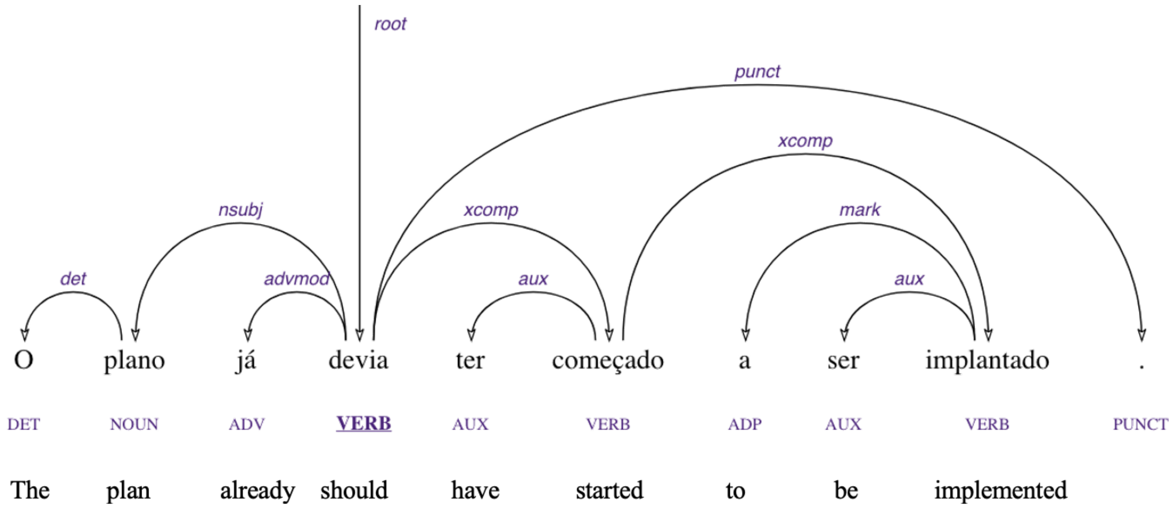
root
punct
xcomp
nsubj
xcomp
mark
det
advmod
aux
aux
O plano já devia ter começado a ser implantado .
DET NOUN ADV VERB AUX VERB ADP AUX VERB PUNCT
The plan already should have started to be implemented

Figure 3: Hybrid annotation of auxiliaries and auxiliated verbs

Figure 4 shows annotation of all verbs as AUX, except for the last one, which is a full verb and takes the root.

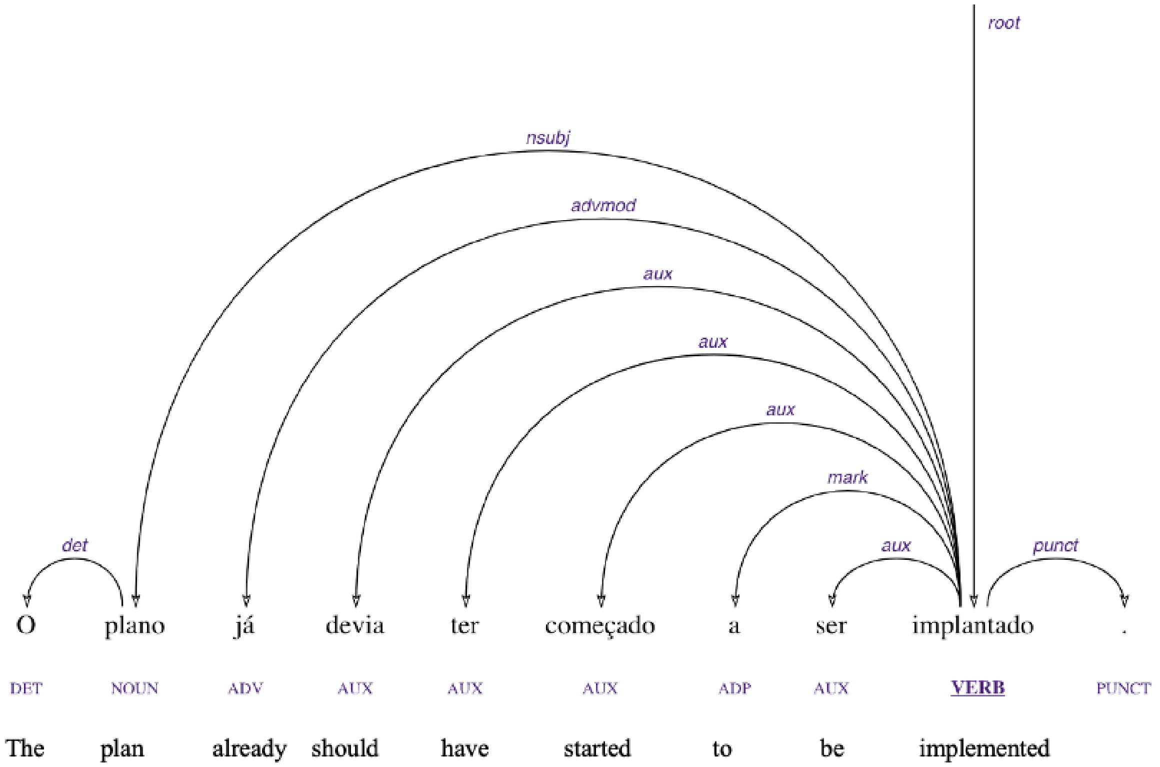
root
nsubj
advmod
aux
aux
aux
mark
det
aux
punct
O plano já devia ter começado a ser implantado .
DET NOUN ADV AUX AUX AUX ADP AUX VERB PUNCT
The plan already should have started to be implemented

Figure 4: Annotation of (auxiliated) full verb as syntactic head

Annotation in Figure 2 is the most satisfactory as it tags all verbs alike, which is a better representation in the case of a chain, where one verb is concomitantly auxiliary to the following verb and auxiliated by the preceding one. Annotation in Figure 3 is less satisfactory; although it keeps the traditional annotation of more grammaticalized auxiliary verbs as AUX, it misses details regarding the syntactic relation between the verbs annotated as VERB and those annotated as AUX, such as the requirement for an infinitive form after the aspectual verb *começado* (*started*). It also splits the verb chain into two xcomp relations. Annotation in Figure 4 is the least satisfactory of all, as it ignores the interdependence relationship between the verbs in a verbal chain and poses a problem regarding the distance between the verb holding inflections and the subject, assuming that the longer the distance is, the more difficult the task of subject detection becomes.

In order to find out which strategy is more suitable for machine learning, Lhoneux et al. (2020) compared two ways of annotating auxiliaries in UD: the auxiliary as head of its auxiliated (head left and dependent right) and the auxiliary as dependent on the auxiliated (head right and dependent left). They concluded that the information the annotation brings to the process depends on the machine learning modeling choices, and, therefore, if properly modeled, the same properties may be automatically acquired. Given that machine learning seems to deal equally well with both forms of annotation, we believe that it is preferable to choose a learning model based on what is desired to be learned rather than to choose a form of annotation based on what the available models can learn.

In Brazilian Portuguese auxiliary constructions, the inflectional head is an auxiliary and, for syntactic purposes, it is the head of the construction, as it must agree with the subject. In UD, however, when a verb is annotated as AUX, it becomes a dependent in a dependency relation **aux**, the head being another verb: a full verb or other auxiliary verb annotated as VERB if there is a chain. Moreover, UD defines AUX as a functional word and restricts its selection as head in a relation. This means that the phenomenon of over-auxiliation (one auxiliary modifying another) cannot be represented using AUX in UD.

From the perspective of semantic applications in NLP, treatment of auxiliaries and full verbs has important implications. Promoting the full verb to head is interesting to information extraction. Buiko et al (2009), for example, compared the effect of different dependency representations on information extraction and concluded that the *trimming* of auxiliary structures enhanced the event extraction results. The trimming of auxiliary structures is an operation that seeks to "prune the auxiliaries/modals as governors from the dependency graph and propagate the dependency relations of these nodes to the main verbs" (Buiko et al 2009). For temporal expression, aspectual and modal verbs are fundamental cues, as explained by Pustejovsky et al (2017) in their design of the TimeML model, aimed to extract time information on events. Modal verbs are also relevant cues for speculation detection, as explored for different domains in Ozgur & Radev (2009), Zhou et al. (2010), Sauri & Pustejovsky (2012), and Rivera Zavala & Martinez (2020).

Since an auxiliation chain has the same subject and refers to a same event, identifying them is productive for extraction tasks, even if auxiliaries are annotated as head or not, and even when there are other functions in between, such as adverbs and pronouns.

Three main arguments are worth summing up at this point:

- there is syntactic relationship between the auxiliaries of modality, tense, aspect and passive voice diathesis within a chain of over-auxiliation;
- verbs in a chain can be at the same time auxiliated by a verb and auxiliary to another one;
- the UD guidelines do not allow an auxiliary verb (considered a functional word) to be head of a dependency relation;

Bearing upon the above arguments, we believe the most productive way to annotate auxiliary verbs in UD is using the tag VERB and relating verbs to each other as open clausal complements (**xcomp**), this relation tag implicating they all have the same subject. Moreover, considering that modality, tense, aspectual and passive voice cues are relevant for many NLP applications, we propose to add a new annotation for this purpose at the morphological level: a feature called *AuxiliaryType*, with the initial values: Tense, Modality, Aspect, and Voice. Thus, regardless of whether the verbs that participate in the auxiliation chain have been annotated as AUX or as VERB, they will be identified at the feature level by the auxiliary function they perform within the chain. The absence of this feature means that the verb is not performing an auxiliary function within the chain and is therefore a full verb.

Our proposal has a twofold impact:

- it allows treebanks with other annotation decisions for auxiliaries to reconcile their annotations by simply adding a feature to indicate auxiliary type (no alteration in annotation needed, but merely addition of features);
- it enables recovery of auxiliation chains, as a sequence of verbs and/or auxiliaries that present a value of AuxiliaryType and are followed by a verb with no value of AuxiliaryType (the full verb, which is the semantic head).

A further proposal is specifying Voice at the feature level. UD provides features to discriminate categories of Modality (Mood), Tense and Aspect. However, it has no feature to discriminate between categories of Voice (values: Passive, Agentive, Resultative, Causative, etc.). It would therefore also be desirable to create a Voice feature to complete the description of auxiliaries in UD morphology. This is particularly interesting because Voice is directly linked to the semantic role of the subject (Patient, Agent, Beneficiary, Cause, etc.), and this would favor other NLP semantic applications.

5 Final Remarks

In this paper, we have discussed the annotation of auxiliary verbs under the UD model, using evidence from linguistic theory to reconcile different ways of annotating phenomena that share semantic similarity, but differ greatly in syntactic behaviour.

As we have argued in the preceding sections drawing on grammaticalization studies, auxiliary verbs are verbs in their own right, i.e., they are neither a closed class of words (they are open to new candidate forms to auxiliaries), nor are they a fully functional class of words (they can operate as full verbs themselves). Throughout this paper, we have put forward arguments in favour of annotating auxiliary verbs as heads in dependency relations whenever the inflected verb is an auxiliary verb, as in Portuguese auxiliary chains. Hence, our proposal is to leverage the role of auxiliary verbs in the syntax of the clause, both in determining the form of auxiliated verbs and establishing agreement with the subject.

We also proposed the inclusion of a new morphological feature, *AuxiliaryTyp*e, with the initial values of Tense, Modality, Aspect, and Voice. This way, we move towards standardizing UD auxiliaries annotation, enhancing comparability between languages. By annotating information on auxiliary function at feature level, we can reconcile our proposal to annotate auxiliaries with that of other languages that may have adopted other annotation strategies, either because their auxiliary verbs are not inflected, or because they have decided to privilege the semantic head in syntactic annotation.

Acknowledgements

Magali Duran and Thiago Pardo are grateful to The Center for Artificial Intelligence of the University of São Paulo (C4AI-http://c4ai.inova.usp.br/), sponsored by IBM and FAPESP (grant#2019/07665-4). Adriana Pagano wishes to thank the National Council for Scientific and Technological Development (CNPq) for grant No. 310630/2017-7.

References

Anderson, Gregory D. S. 2006. *Auxiliary Constructions*. Oxford University Press.

Baptista J., Mamede N., Gomes F. 2010. Auxiliary Verbs and Verbal Chains in European Portuguese. In: Pardo T.A.S., Branco A., Klautau A., Vieira R., de Lima V.L.S., eds. Computational Processing of the Portuguese Language. PROPOR 2010. *Lecture Notes in Computer Science*, vol 6001. Heidelberg: Springer. https://doi.org/10.1007/978-3-642-12320-7_14

Benveniste, Émile. 2006. Estrutura das Relações de Auxiliaridade. In: *Problemas de Linguística Geral II*. Campinas: Pontes, [1965], pp. 181-198.

Buyko, E., Faessler, E., Wermter, J., & Hahn, U. (2011). Syntactic simplification and semantic enrichment-trimming dependency graphs for event extraction. *Computational Intelligence*, 27(4), 610–644. doi:10.1111/j.1467-8640.2011.0

Heine, Bernd. 1993. *Auxiliaries: Cognitive Forces and Grammaticalization*. New York: Oxford University Press.

Ide, Nancy. 2017. Introduction: The Handbook of LinguisticAnnotation. In: Ide, Nancy; Pustejovsky, James, eds. *Handbook of Linguistic Annotation*. Springer. DOI 10.1007/978-94-024-0881-2_1

Ilari, Rodolfo, Basso, Mario. 2014. O verbo. In Ilari, Rodolfo. *Gramática do português culto falado no Brasil* - vol. III - palavras de classes abertas.

Krug, Manfred. 2012. Auxiliaries and grammaticalization. In: Heine, Bernd; Narrog, Heiko, eds. *The Oxford Handbook of Grammaticalization.* Oxford University Press. DOI: 10.1093/oxfordhb/9780199586783.013.0044

Kuteva, Tania. 2001. *Auxiliation: an enquiry into the nature of grammaticalization*. New York & Oxford: Oxford University Press.

Lhoneux, Miryam de, Sara Stymne, Joakim Nivre. 2020. What Should/Do/Can LSTMs Learn When Parsing Auxiliary Verb Constructions? *Computational Linguistics* 46(4), pp. 763-784.

Lobato, Lúcia M.P. 1975. A auxiliaridade em português. In: Lobato, Lúcia et al., ed. *Análises linguísticas*, pp. 27-91. Petrópolis: Vozes.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal Dependencies v2: An ever growing multilingual treebank collection. In: *Proceedings of the12th Language Resources and Evaluation Conference*, 4034-4043. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.497.

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. In: Alexander Gelbukh, ed. *Computational Linguistics and Intelligent Text Processing*, pp. 3–16. Cairo, Egypt: Springer International Publishing. doi:10.1007/978-3-319-18111-01.

Osborne, Timothy and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics* 4(1):17. pp. 1–28, DOI: https://doi.org/10.5334/gjgl.537

Ozgur, Arzucan, Dragomir R. Radev. 2009. Detecting Speculations and their Scopes in Scientific Text. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1398–1407, Singapore.

Pontes, Eunice. 1973. *Verbos Auxiliares em Português*. Rio de Janeiro, Ed. Vozes.

Pustejovsky, James; Bunt, Harry; Annie. Zaenen. 2017. DesigningAnnotation Schemes: From Theory to Model. In: Ide, Nancy; Pustejovsky, James, eds. *Handbook of Linguistic Annotation*. Springer. DOI 10.1007/978-94-024-0881-2_1

Ridley, Matt. 2010. *The Rational Optimist: How Prosperity Evolves*. HarperCollins Publishers.

Rivera Zavala R, Martinez P. 2020. The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study. *JMIR Med Inform* 8(12):e18953 URL: https://medinform.jmir.org/2020/12/e18953 DOI: 10.2196/18953

Sauri, Roser, Pustejovsky, James. 2010. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*, 38, 2, pp. 261-299.

Steele, Susan. 1994. Reviewed Work: Auxiliaries: Cognitive Forces and Grammaticalization by Bernd Heine In: *Language*, Vol. 70, No. 4, pp. 818-821. Linguistic Society of America Language https://doi.org/10.2307/416332. Accessed August 13, 2021.

Tesnière, Lucien. 1959. *Elements of Structural Syntax*. Osborne, Timothy; Kahane, Sylvain, translators. John Benjamins Publishing Company, 2015.

Zhou, H., Li, X., Huang, D., Li, Z., & Yang, Y. 2010. Exploiting Multi-Features to Detect Hedges and their Scope in Biomedical Texts. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task.* pp. 106–113. Uppsala: Association of Computational Linguistics.