A Universal Classification of Lexical Categories and Grammatical Distinctions for Lexicographic and Processing Purposes

Roser Saurí, Ashleigh Alderslade, Richard Shapiro
Oxford University Press
E-mail: roser.sauri@oup.com, ashleigh.alderslade@oup.com, richard.shapiro@oup.com

Abstract

We introduce COMO (Compositional Morphosyntactic Ontology), a classification of part-of-speech categories and their associated grammatical features, which aims to be valid across languages of very different typology. The work has been carried out within the context of the Oxford Global Languages programme, which has the goal of developing language knowledge for 100 languages, particularly those under-represented in the digital space. The requirements around this project are: to be able to describe languages of different typeS while respecting their grammatical tradition, and to be able to serve two main use cases that define our typical work, namely, the labelling of linguistic information in lexicographic products, and the provision of support for language processing tools and corpus annotation processes. These requirements determined the conception and design of COMO, created as a reference model within a broader data architecture in order to address issues of syntactic and semantic interoperability. Our proposal builds on top of previous initiatives in the field aiming at the same goals, but incorporates different features in order to accommodate for the requirements in the project.

Keywords: part-of-speech tagging, morphosyntactic information, language modelling, interoperability, multilinguality

1 Introduction

The codification of each language (dictionary and grammar) has traditionally been carried out according to distinctions identified within the language, or at best in line with distinctions already used in similar ones. For example, the concepts and terms employed to account for the tense system in different Romance languages have an indisputable resemblance, but differ in some respects to the treatment of the same system in English. At the part-of-speech (POS) level, differences are also evident. Some linguistic descriptions present lexical categories specific to the language they describe (e.g., phrasal verb for English), or use different terms to refer to the same kind of units (e.g., demonstratives functioning as noun modifiers can be treated as either determiners or adjectives, depending on the linguistic tradition of the language).

This is acceptable when working within the scope of a dictionary or grammar of a particular language, and even in the case of bilingual lexicography. However, the approach soon becomes insufficient for modern multilingual usages, because there is no common ground that allows the different language descriptions to meet, thus providing little opportunity for interoperability among resources or applications of different provenance.

The problem of heterogeneous linguistic encodings is not new. It was already on the agendas of EU-funded projects in the early 1990s, when the Expert Advisory Group for Language Engineering Standards (EAGLES) was created in order to provide guidelines and standards for encoding and managing linguistic resources in different languages. At that moment, the effort was mostly

focused at ensuring *syntactic interoperability* among systems (dictionaries, corpora, etc.) of diverse origin and languages; that is, at ensuring that they could share information or communicate given a common data model and format. Syntactic interoperability, however, does not guarantee that the interpretation of the information shared by the systems is the same (Ide & Pustejovsky 2010). An example of this in the area of morphosyntactic information is brought by the term *absolute*, which in some English dictionaries is applied to pronouns like *yours* and *theirs*, to express that they only have a pronominal use (as opposed to e.g. *her*), whereas in Romanian it refers to transitive verbs when used without an object.

The limitations of systems not fully communicating at the semantic level have become apparent in more recent times, when the digitization of linguistic knowledge and the development of natural language processing (NLP) tools have started to expand to languages of very different typological affiliation (in the last decade, Chinese, Japanese, and Arabic; more recently, languages from the Indian subcontinent). Adding to this, there have also been advances in the technology supporting such digitization, which now allows for a fully-linked data paradigm. In the linguistics field this has materialized into the Linguistic Linked Data program (LLD, Chiarcos, McCrae et al. 2013), which makes it possible to gather linguistic datasets of very different languages in the same repository, connect these at a lexical level, and elicit additional knowledge by automatic means.

To allow for full communication between different linguistic resources, a number of more recent projects have been deployed specifically targeting *semantic interoperability*. All of them rely on the existence of a reference model where domain concepts are unambiguously defined, and to which the different linguistic resources defer in order to communicate amongst themselves and interpret information meaningfully. Some of these projects are: the General Ontology of Linguistic Descriptions (GOLD), the ISO TC37/SC4 Data Category Registry (ISOcat), the set of Ontologies of Linguistic Annotation (OLIA), the Universal Dependencies project (UD), and the Universal Morphological Annotation (UniMorph).[1]

The challenges addressed by these projects are also shared at the Dictionaries Division of Oxford University Press (OUP), specifically within the context of the Oxford Global Languages (OGL) program.[2] OGL aims to develop linguistic knowledge for 100 languages with a particular focus on those which are under-represented in the digital sphere. This includes languages with very little codification or with features quite different from Indo-European ones, which are typically the languages that have inspired the grammatical distinctions commonly assumed in linguistic analysis. Furthermore, at the technical level the project plans to store and integrate the information for the resources of these many languages in a centralizing repository, along the lines of the Semantic Web paradigm. A key factor in this is having a model able to accommodate complexities across these many languages and enable the linking between them (Parvizi et al. 2016).

The current paper focuses on the fragment of the model concerning morphosyntactic information, that is, POS classes (aka lexical categories) and grammatical distinctions manifested by morphological and syntactic means. In particular, it presents COMO (Compositional Morphosyntactic Ontology), which aims to be a universally valid classification for distinctions at this linguistic level. We are aware that there is disagreement on whether it is possible to set a universal POS classification (Evans and Levinson 2009). However, we also appreciate that there is a set of coarse POS classes that can be commonly found across many languages.

Some of the projects listed above already present a reference model for that information level, and therefore have served as very valuable starting points for our work. However, the requirements

1 Full references for each of these projects will be provided in Section 4.

2 https://www.oxforddictionaries.com/ogl

imposed by the nature of our project and the work activity around it make them not fully useful for our purposes. The paper will thus introduce our classification by first reviewing the project requirements and subsequent design choices, and then comparing these against the morphosyntactic classifications that are most relevant here.

2 Requirements and System Design

Our proposal is based on a set of requirements imposed by the wide linguistic scope of the OGL program, the nature of activities and resources managed in the Dictionaries Division at OUP, and the subsequent use cases that we need to serve. These requirements determined our approach and shaped the design of the universal classification put forward in this paper. They are presented in the following subsections.

2.1 Cross-linguistic Validity

An OGL classification of morphosyntactic information must be able to serve across languages of very different typologies, which therefore diverge greatly in how they encode grammatical distinctions. A feature may be expressed in some languages via morphological mechanisms (for example, verbal tense in Romance languages), encoded in others using independent particles, and in other languages not expressed at all (e.g., Chinese). As a result, dictionary-based classifications of POS categories and associated grammatical information tend to be modeled on the language (or languages) targeted in each individual project. By contrast, a universal classification needs to be able to reflect the commonalities across languages in spite of the diverse means of organizing information through grammar in each language system. Two situations can be distinguished here.

2.1.1 Considerations on distinctions at the POS category level

Grammatical distinctions are in some languages manifested at the morphological level, while in others are expressed by means of independent lexical units. An example of that involving English concerns the grammatical feature of verbal infinitive. Whereas in many languages this is realized as part of the verb morphology, in English it is expressed by means of the marker *to*. As a result, it is not uncommon that languages in which elements like these are independent lexical units classify them with an ad hoc POS category not applicable to other languages. Such POS categories therefore lack universal validity.

This problem can be avoided if POS categories are defined appealing to basic configurational properties. For instance, syntactically, what do they combine with to form more complex units (or phrases), and what are the roles that they play in the grammatical organization of the sentence; morphologically, what type of inflection process, if any, do they undergo; etc. From this perspective, independent lexical units dedicated only to encoding grammatical information can be classified as POS categories of cross-lingual validity, most often of closed-class type, such as particles or adpositions. For example, the marker for concord, commonly considered as an independent POS in Bantu languages, can actually be classified as belonging to the more general category of affix.

2.1.2 Considerations on morphosyntactic distinctions within POS categories

Many POS classifications tend to subclassify these based on the morphosyntactic distinctions that seem more relevant in each language. Taking verbs as example, we see that in the English tradition, a common way of subclassifying them is distinguishing between lexical and auxiliary verbs. By

contrast in Russian, where aspect distinctions are fully lexicalised, verbs are subclassified into perfective and imperfective, while in the tradition of some Romance languages the direct subdivision of verbs is set in terms of their subcategorization structure (transitive, intransitive, etc.).

Nevertheless, a system where POS classes are tied to particular grammatical features will fail the purpose of being valid across languages. First, each POS will have to subdivide into as many sub-classifications as are found across the different languages covered. This situation will get even less manageable considering that for some POS categories, there may be several levels of information that apply. For example in Spanish, verbs can be simultaneously featured according to their subcategorization and pronominal properties, while in English, nouns are categorized based on their type (common, proper) and countability properties (mass, countable).

Second, the system will introduce a lot of redundancy given that many of the grammatical distinctions are shared across POS classes. For instance, degree features (positive, comparative, superlative) can be found in adjectives and adverbs. Similarly, in many languages the distinctions used in the classification of determiners (e.g., demonstrative, exclamative, indefinite, interrogative, possessive) also apply to pronouns.

Therefore, a universal POS classification that includes all POS classes and subclasses independently set for each language is not viable for our purposes. Instead, we propose a system that represents the grammatical distinctions that are possible in any language by means of a set of features complementary (and therefore orthogonal) to the basic POS ontology. Each feature is modelled as an attribute with its respective set of values, e.g., attribute *number* has as possible values: *singular, plural, dual, trial, invariable*, etc.

This results in a highly compositional approach in which the list of POS categories is kept to a minimum number of distinctions, as universally valid as possible and therefore general enough to serve languages of very different typologies. Furthermore, any POS class can in addition be qualified with several attribute-value pairs of grammatical distinctions. For example, a Spanish '*impersonal transitive verb'* will bear the pairs *pronominal_type:impersonal, subcategorization: transitive.*

2.2 Respectful of the Grammatical Tradition In Each Language

Each language is described by its own grammatical tradition, as reflected in the way it is presented and taught in grammar books, dictionaries, etc. A major target of our work here is precisely producing and publishing dictionaries for different languages, and therefore the linguistic classifications used should be in agreement with those commonly assumed in the grammatical tradition of each language.

However, grammatical classifications in each tradition tend to be constrained to the language they describe, therefore precluding a wider, cross-linguistic view of grammar distinctions, which is what is aimed here. Two additional issues resulting from adopting the terminology in each language tradition are *feature redundancy* and *concept conflation.* The former occurs when a classification has different terms for the same notion, as the result of inheriting the terms and definitions from different language descriptions. The latter refers to a situation in which the same term is used for two different concepts, due to the fact that different languages' descriptions use it in different ways (e.g., the example with the term *absolute* presented above).

Overall, these issues derive from a wider problem, namely, that of heterogeneous linguistic descriptions constraining the reusability and interoperability of linguistic datasets and resources, which has led to a quite intense area of work in the last decade (see, e.g., Chiarcos & Erjavec 2011; Ide at al. 2017). Two solutions have been put forward in order to enhance the consistency of linguistic categorizations across languages:

- Using *cross-linguistic meta schemes* that include a fix set of content categories to be adopted by all languages. This is the approach adopted by the early initiatives in the field, such as EAGLES (Leech & Wilson 1996) and MULTEXT (Erjavec 2010, 2012).
- Providing a *reference terminology as interlingua* between the different linguistic encodings. Terms in this reference model must be defined so there is no ambiguity in their usage, and to avoid feature redundancy and concept conflation. In addition, a set of linking specifications must be developed for each language-specific morphosyntactic classification, to ensure proper mappings between each resource and the interlingua terminology. Linguistic terminologies such as GOLD (Farrar & Langendoen 2003), ISOcat (Kemps-Snijders et al. 2009), and OLIA (Chiarcos & Sukhareva 2015) assume this approach.

We adopt the second solution: COMO is conceived as a reference model facilitating the harmonization of terms and distinctions used in language-specific resources and applications, which are then able to maintain their original model and yet communicate via a set of mapping models.

2.3 Able to Serve Different Use Cases

COMO must be able to support different use cases: from the encoding of dictionary content to the annotation of corpus data, passing through the codification of the NLP tools used for the automatic processing of language. Because of this, the morphosyntactic classification must be able to account not only for the prototypical POS classes and their grammatical distinctions, but also for other kinds of units.

Corpus content and NLP tools, for example, require sensitivity to punctuation marks or non-standard lexical elements, such as symbols. We decided to set a category for these at the same level as more standard POS categories.

Similarly, the lexicographical use cases of OGL require the inclusion of other elements not typically considered to be POS categories. These are: parts of lexical units (i.e., affixes) as well as aggregates of lexical units, such as contractions or different types of multiword expressions (phrases, idioms, etc.).

3 The Compositional Morphosyntactic Ontology (COMO)

The Compositional Morphosyntactic Ontology (COMO) is a repository of linguistic terminology that provides common ground for languages of very different type, which historically have been described through diverse grammatical traditions. It enables harmonization of linguistic annotations in different types of resources, such as lexicographical datasets (including machine-readable dictionaries and dictionaries for human consumption), text corpora, and language processing tools.

COMO is not conceived as a cross-lingual meta-scheme to be assumed by all languages and resources. Rather it is an interlingua, a reference model to which morphosyntactic annotations in different resources and languages map in order to allow for maximum interoperability.

Because of this, it was vital that no language was a stronger driver than any other in the setting of a POS classification.

COMO defines and organizes morphosyntactic concepts in an ontological structure by appealing to criteria and definitions of cross-lingual validity as much as possible.

3.1 Lexical Categories

Table 1 presents the full set of lexical categories (aka POS classes) in COMO. It provides comments defining the category only when deemed necessary.

Table 1: POS classes and subclasses in COMO

Lexical category	Comments	
adjective		
adposition	Cover term for a closed class of words that express spatial or temporal relations, or mark various semantic roles. Typically combining with one complement, generally a noun phrase. Divided into the following three subclasses based on the position they take with respect to the complement: before (preposition), after (postposition) or surrounding it (circumposition). Possible subclasses are:	
	circumposition	Consisting of two parts that appear on each side of the complement.
	postposition	Following the complement.
	preposition	Preceding the complement, as in English.
adverb		
affix	Morpheme attached to a stem to form a new word. In some cases it is written as part of the same word whereas in others it appears as an independent element.	
	circumfix	Two separated parts appearing on each part of the stem.
	combining_form	Word normally used in compounds in combination with another element to form a word (e.g. *Anglo-* 'English' in *Anglo-Irish*).
	infix	Appearing within the stem.
	prefix	Appearing before the stem.
	suffix	Appearing after the stem.
article		
conjunction		
contraction	Combination of two or more words belonging to a different POS into a single lexical unit. For example, the combinations of preposition + article in French: *du* (*de+le*).	
determiner		
ideophone	Ideophones are lexical units that evoke a vivid impression of certain sensations or sensory perceptions (e.g. sound meow for a cat), movement, colour (e.g., English *bling*, describing the glinting of light on things like gold), shape, action (*ta-da!*), etc. It is a lexical class based on the special relation between form and meaning. In some languages, ideophones correspond to common POS classes (e.g., adjectives, adverbs, etc.), but in others they are an independent POS.	
idiomatic	Multiword, phrasal or clausal expressions, generally with no compositional interpretation.	
interjection		
noun		
numeral		
particle	Particles must be associated with another word or phrase to impart meaning. They typically encode grammatical distinctions like negation, mood, tense, or case), etc. However, they cannot be classified as other main POS, including functional ones, such as prepositions, conjunctions, etc.	
predeterminer		
pronoun		
punctuation	left_parenth_punc	Left parenthetical punctuation mark, e.g., (, [.
	right_parenth_punc	Right parenthetical punctuation mark, e.g.,), }
	sentence_final	Sentence final punctuation mark
	sentence_medial	Sentence medial punctuation mark
residual	Cover class for non-standard forms, such as symbols or digits.	
verb		

POS classes have been determined according to the basic configuration (grammatical and syntactic) properties of the elements being classified. For example, syntactically, what does a lexical element combine with in order to form a more complex unit, or what is the role that it plays in the grammatical organization of the sentence; morphologically, what type of inflection process, if any, does it allow? And so on.

Since POS is essentially a configurational distinction, we have established subclasses only if they respond to strictly positional criteria that can be applied across languages. In particular, the classes *adposition*, *affix*, and *punctuation* have been subdivided based the possible placements of their elements. Note that these subclassifications are different from those in other reference models, such as GOLD or OLIA, which use grammatical distinctions to subclassify POS categories (e.g., *transitive verb* or *demonstrative determiner*). The classification also includes classes for material not traditionally considered part of the grammar, such as punctuation marks (class *punctuation*), or symbols (under class *residual*).

3.2 Grammatical Features

Complementing the POS classification, there is a set of features covering the different grammatical distinctions that can be manifested in a language, from those most commonly found in Indo-European languages (like tense, case, number, and gender) to others associated with languages which have less coverage in terms of linguistic encoding and language resources (for example, concord for Bantu languages, or classifier type for Chinese and other languages). We refer to these features as the set of *grammatical features*.

Previous proposals of morphosyntactic classifications also account for grammatical distinctions. EAGLES (Leech and Wilson 1996) and MULTEXT (Erjavec 2010, 2012) are designed as *positional tagsets*, that is, as sets of tags in which each tag is modelled as an acronym, and where each piece of information (POS class and the possible grammatical features for that POS class) is represented in a specific position of the tag with a one-character code identifying the corresponding value. For example, in EAGLES, a common noun, feminine singular and in accusative case would be expressed as a tag like: *NCFSA*. The problem with this approach, however, is that it imposes a rigid structure. This makes it difficult to add new features as languages with yet uncovered properties (e.g., noun class for some Bantu languages) are added to the repository.

Other morphosyntactic classifications have adopted a more flexible approach by distinguishing between POS categories on the one hand, and morphosyntactic features on the other; e.g., GOLD (Farrar & Langendoen 2003) and OLIA (Chiarcos & Sukhareva 2015). Nevertheless, they still present at least one level of POS subcategory that tends to be language-specific, and that leads to the issues identified in section 2.1.2.

By contrast to all this previous work, COMO adopts a fully compositional approach, along the same lines as the Universal Dependencies (UD)[3] and the Universal Morphological Annotation (UniMorph)[4] proposals, where grammatical distinctions of all kinds are represented as information independent from the POS categories. In particular, each feature is modelled as an attribute with its corresponding set of values.

A key element in this approach is the level of granularity of the feature. It is important to separate into different features grammatical information that may manifest simultaneously but in fact corresponds to different notions, as some languages may present one feature but not the other. For instance, event

3 http://universaldependencies.org/ [25/03/2018]

4 http://unimorph.org/ [25/03/2018]

duration (distinctions of punctual vs. non-punctual) and telicity (distinctions of telic vs. atelic) should be considered as independent from each other, since they can combine in different ways.

There is no imposed hierarchy of one feature over another, thus avoiding conflicts between language traditions. Moreover, new features and values can be added as further languages are included in the OGL project, without having an impact on the previously described languages.

The rich morphosyntactic branch in OLIA was a solid base to model this kind of information. However, some of the languages that we intend to model have complex morphosyntactic features, such as Northern Sotho, wherein a single orthographic word may contain a number of morphemes; others have extensive noun systems, such as isiZulu, which has 17 different classes. OLIA was unable to fully accommodate modelling of such features.

Our proposal was also informed with the lexical and grammatical labels used in OUP monolingual and bilingual dictionaries. Finally, a further source of information was other well-tested classifications for morphosyntactic knowledge developed with multiple languages in mind or as a collaborative effort among teams in different countries (EAGLES, MULTEXT).

At the moment, COMO presents 46 grammatical features, which range from the basic nominal, verbal and adjectival morphological features present in many languages (e.g., *number, gender, case, degree, person, mood, tense*, etc.) to elements codifying syntax (e.g., *subcategorization patterns*), distinctions at the lexical semantic level (e.g., *aspect, telicity, countability*), or pragmatic information (e.g., *definiteness, referentiality, evidentiality, sentence modality*). Some of these categorizations are present in multiple POS classes (*degree*, *number*, *gender*), whereas others are particular to only one (*voice*, *pronoun function*). Furthermore, most categorizations are shared across several languages, although a few cases had to be tailored to specific ones, such as the classifications on *diptoticity* or *verb form type* for Arabic.

4 Comparison with Other Morphosyntactic Classifications

The last few decades have seen a significant effort, especially from the NLP community, to develop universal classifications for the description of different levels of linguistic information (lexical, morphological, syntactic). Early work in this respect includes initiatives on corpus annotation, like the EAGLES guidelines for annotating morphosyntactic information in corpora and lexicons (Leech & Wilson 1996), which led to the development of POS tagsets such as FreeLing[5] (Carreras et al. 2004) and MULTEXT-East (Erjavec 2010), among others.

This section describes the most relevant morphosyntactic classifications aiming to address interoperability issues at the level of linguistic information, and assesses them against the requirements introduced in Section 2.

4.1 MULTEXT-East

The MULTEXT-East morphosyntactic specifications[6] (Erjavec 2010, 2012) provide attributes and values for annotating at the word level, focusing in particular on 16 Eastern European languages. This system distinguishes 14 main POS classes (called morphosyntactic categories in the proposal), each of them possibly splitting into several levels of subcategories. Considering all possible categories and subcategories, there are a total of 127 morphosyntactic classes. This approach has the issues

5 https://talp-upc.gitbooks.io/freeling-4-0-user-manual/content/tagsets.html [25/03/2018]

6 http://nl.ijs.si/ME/owl/ [25/03/2018]

and risks presented in Section 2.1.2 with regard to cross-linguistic validity. In fact, some distinctions seem quite clearly based on the grammatical traditions of different languages. For example, the differentiation between *DemonstrativeQuantifier* and *DemonstrativeDeterminer*, respectively under the *Quantifier* and *Determiner* categories, is not clear. The same can be observed with regard to the grammatical features complementing the POS classification. For example, the description of the feature *SyntacticType* is different depending on the language it is used for. In general, MULTEXT-East seems unable to successfully handle languages others than those for which it was developed.

4.2 General Ontology of Linguistic Descriptions (GOLD)

GOLD[7] (Farrar & Langendoen 2003) is a very complete ontology accounting for information at several levels of linguistic description, from phonetic to morphosemantic properties, and covering also structural unit or human language variety. The layers that are of interest here are *POS Properties* and *Morphosyntactic Properties*. POS classes are grounded on very strong language-agnostic principles that focus on the function of the item in the sentence. As a result, what are usual terms in our linguistic tradition, such as *verb*, *conjunction*, or *pronoun*, are nested under more general terms such as *Predicator*, *Functor*, or *Pro Form*, respectively. There are 19 main POS classes, which can then split into several layers of subclasses, leading to a total of 91 classes.

GOLD offers a very interesting perspective on POS classification because of its language-independent approach. It is a terminology repository very comparable to COMO in purpose, scope, and technical approach. First, it was created to address the problem of having different markup schemes for annotating linguistic data, and therefore to provide a unified description of data in different languages. More specifically, it originated to handle the annotation and description of endangered languages, a purpose very close to OGL's main goals. Furthermore, it seeks to be compatible with Semantic Web approaches, and therefore to enable automated reasoning over linguistic data.

There are, however, several reasons that prevented us from adopting it. Firstly, its linguistic terminology is quite different from that more commonly used in dictionaries and other language resources. Secondly, although it contains most of the more standard concepts in our tradition (e.g., verb, adjective, conjunction), it organizes them in a structure more complex than what is actually needed for our purposes. For example, the class for subordinating conjunctions is in level 3 of the embedding, within *Connective* within *Functor*. Finally, in spite of its wide coverage of linguistic phenomena, it does not account for categories that are important when dealing with text annotation and language processing tools, such as punctuation marks and symbols.

4.3 ISO TC37/SC4 Data Category Registry (ISOcat)

ISOcat (Kemps-Snijders et al. 2009) is a web-based repository of linguistic terminology providing uniform naming and semantic descriptions in order to facilitate interoperability of a wide range of resource and application types.

It took a community-driven approach from quite early in its development, allowing everybody to extend the repository based on specific languages or project needs. That approach led to the proliferation of data categories, and caused issues of redundancy (different terms defining the same notion) or concept conflation (the same term for two or more notions). See Ide, Calzolari et al. (2017:136—139) for a more comprehensive description of the situation.

With regard to the requirements of our project, these problems translate into the presence of POS categories specific for only one language, and a poor handling of the top level classification for POS

7 http://www.linguistics-ontology.org/gold/2008 [25/03/2018]

tags, with categories that can be grouped under more general classes (e.g., the different punctuation mark signs). In addition, ISOcat also includes subcategories as part of the POS classification, which leads to more than 100 POS classes and therefore carries the issues identified earlier with this type of approach (refer to section 2.1.2).

4.4 Ontologies of Linguistic Annotation (OLIA)

OLIA[8] (Chiarcos & Sukhareva 2015) was developed to enable semantic interoperability among linguistic resources of diverse types and annotated at different levels of analysis, i.e., from phonology all the way up to discourse structure. Unlike other initiatives that aim at the same purpose (e.g., GOLD, ISOcat), OLIA is a complete architecture that includes: (a) a *reference model* with terms and their definitions, to be used as interlingua for mapping linguistic resources of different provenance and tradition; (b) specific *annotation models* for different levels of linguistic description (morphology, morphosyntax, syntax, discourse), consisting of annotation schemes and tagsets for more than 85 languages; (c) a set of *linking models* mapping concepts and properties in each of these annotation schemes to the reference terminology; and finally also (d) a set of *linking models for mapping external annotation models* to the OLIA reference model (as is the case for MULTEXT-East, GOLD, ISOcat, or Universal Dependencies). Moreover, OLIA ontologies can serve as a centralizing hub for linguistic data categories within the Linguistic Linked Open Data (LLOD) framework[9] (Chiarcos et al. 2013).

OLIA is a very rich and powerful resource that already serves two of our requirements in Section 2. First, given its purpose of serving different types of linguistic resources and applications, the OLIA reference model can address the use cases we presented in Section 2.3. That is, it accounts for categories that go beyond standard POS classes but that are needed for NLP applications (e.g., punctuation marks, symbols) or lexicographic products (contractions, multiword expressions). Second, its four-tier architecture allows us to be respectful of the grammatical tradition of each language or the annotation schemes already adopted in different corpora and NLP tools (a requirement in Section 2.2).

However, OLIA cannot equally well serve our need for a classification of POS categories and grammatical features with cross-linguistic validity (a requirement in section 2.1). The POS classification in its reference model follows the standard approach of splitting POS categories into subcategories depending on the features that are considered more prominent for that category in each language. This leads to an enumerative model, where, for example, a common noun belongs to a different class (*CommonNoun*) than a mass noun (*MassNoun*), and thus the linking model must then take care of cases in which the two concepts apply to the same linguistic unit.

This enumerative approach also results in some unclear areas in the classification, especially regarding elements of lesser presence in Indo-European languages. There is for example the *Unique* class, which is conceived in accordance with the definition for this same category in EAGLES, as approximating "the linguistic concept *Particle*. It covers categories with unique or very small membership (…)".[10] In OLIA, this class is split into almost 30 subcategories that would be handled more coherently and systematically using a higher compositional approach, like the one adopted in COMO. *Particle* is a key category in any morphosyntactic classification aspiring at universal validity.

8 http://acoli.cs.uni-frankfurt.de/resources/olia/ [25/03/2018]

9 http://www.acoli.informatik.uni-frankfurt.de/resources/llod/ [25/03/2018]

10 http://www.ilc.cnr.it/EAGLES96/annotate/node16.html#mp [25/03/2018]

4.5 Universal Dependencies (UD)

UD[11] is a community-based project that has the goal of developing grammatical annotations consistent across languages, with the purpose of enabling the development of NLP tools, and therefore annotated corpora, from a language typology perspective. In particular, UD provides tagsets for syntactic dependencies, POS categories, and morphosyntactic features, respectively based on the set of Stanford dependencies (de Marneffe et al. 2008, 2014), Google universal POS tags (Petrov et al. 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman 2008).

Specifically concerning POS tags in UD, the classification results from an attempt to simplify the quite fine-grained categories employed in most treebanks, after some analysis and experiments suggesting that coarser POS classes could help in the NLP tasks of POS taggers and grammar induction. Resulting from this, UD defines 12 basic POS categories and a set of grammatical attributes with their possible values. In addition, it provides mappings for a number of treebanks, accounting for more than 60 languages of very different typologies.

Our work shares the foundations and main guidelines of the UD project: cross-linguistic validity and suitability for consistent tagging in corpora and NLP tools, disregarding the grammar tradition assumed in the resource. Consequently, there are the following key common elements between the two projects:

- The interlingua architecture approach, with a terminological repository set as reference model across languages, and then a set of linking models for mapping the annotations in different resources and traditions to the concepts in the reference model.
- The compositional approach to morphosyntactic categorization, that is, by means of a set of basic cross-linguistically valid POS classes, which is further complemented with a set of grammatical features.
- A minimum set of coarse POS classes.

The main difference between UD morphosyntactic classification and COMO, however, has to do with the requirement of being able to encode content present in dictionary products, such as parts of words (affixes) and word aggregates (contractions, idioms, etc.).

4.6 Universal Morphological Feature Schema (UniMorph)

Similar to UD, UniMorph[12] is a collaborative project developing a cross-linguistically valid morphological classification to be used for training language processing tools. UniMorph aims to overcome the problems of previous classifications that are strongly driven by one, or a particular set of, languages. Thus, it is conceived as an interlingua that allows for syntactic and semantic interoperability among linguistic resources.

UniMorph does not provide POS categories. It focusses only on inflectional morphology categories that represent dimensions of meaning (e.g., number, gender, case, degree). It includes 23 dimensions (equivalent to our grammatical features), which are further specified by means of one or more features (corresponding to our values) from a set of 212 items (Sylak-Glassman 2016).

UniMorph has important similarities with our model, such as the concern for an interlingua capable of avoiding the issues of feature redundancy and concept conflation, the strongly hand-engineered approach to ensure this, and an emphasis on identifying what can be considered the semantic atoms (that is, the meaning elements that cannot be decomposed into a more fine-grained interpretation units) in order to guarantee a compositional approach.

11 http://universaldependencies.org/ [25/03/2018]

12 http://unimorph.org/ [25/03/2018]

Nevertheless, it also differs significantly from our proposal. First, it does not provide a classification of POS categories that can be considered valid across languages. Second, it is built top-down by reviewing the language typology literature and gathering the linguistic traits that are described for the different languages, whereas our approach is bottom-up, accounting for what it is brought in by the languages under consideration, and adapting the system as needed. Finally, it includes only distinctions introduced by inflectional morphology, whereas COMO accounts also for traits pertinent to other levels of the linguistic description as long as they are encoded as part of the information in lexical units. Examples of this include subcategorization type (e.g., transitive, intransitive, bitransitive, etc.), numeral type (cardinal, fraction, multiplier, ordinal, etc.), or pronoun scope (exclusive, inclusive).

5 Status and Future Work

We have introduced COMO, a classification of POS categories and their associated grammatical features, which aims to be valid across languages of very different typologies. POS classes are defined according to basic configurational properties, leaving all grammatical distinctions to be modelled by a set of complementary features, which are handled compositionally and do not bear any hierarchical organization, contrary to most previous proposals. COMO is conceived as a reference model within a wider data architecture, and therefore acts as interlingua that guarantees semantic interoperability while preserving the original encoding of linguistic resources.

Currently, COMO is deployed in RDF and JSON formats, and some language resources are modelled according to that. It is used to support lexical content relating to 20 languages of disparate families, including Indo-European (e.g. English, Spanish, Hindi, Urdu), Austronesian (Malay, Indonesian), Bantu (Swahili, isiXhosa, Setswana), and Quechuan (Southern Quechua).

COMO is still under development, as new languages are added to the OGL program. However, it already serves a wide range of morphosyntactically manifested phenomena. In fact, the classification has been actively used from its inception, and organically tested and enriched with the inclusion of new languages in the program.

In the near future COMO will be made available to the general public. Additional further work involves formalizing the mapping models between the different lexical resources and COMO reference model, as well as developing and deploying a URI strategy that allows this content to be moved into a linked data paradigm. Longer term, we plan to also generate the necessary mapping models to connect our content to external models (e.g., UD, UniMorph) and lexical resources.

References

Carreras, X., I. Chao, L. Padró, M. Padró (2004) FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the Language Resources and Evaluation Conference LREC 2014*: 239-242.

Chiarcos, C., and T. Erjavec (2011) OWL/DL formalization of the MULTEXT-EAST morphosyntactic specifications. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*: 11—20.

Chiarcos, C., J. McCrae, P. Cimiano, C. Fellbaum (2013) Towards Open Data for Linguistics: Linguistic Linked Data. In A. Oltramari et al. (eds.) *New Trends of Research in Ontologies and Lexical Resources.* Theory and Applications of Natural language Processing. Springer-Verlag, Berlin Heidelberg.

Chiarcos, C., and M. Sukhareva (2015) OLiA – Ontologies of Linguistic Annotation. In *Semantic Web*, vol. 6(4): 379-386.

de Marneffe, M-C., T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C.D. Manning (2014) Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of Language Resources and Evaluation Conference 2014*.

de Marneffe, M-C. and C.D. Manning (2008) The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.*

Erjavec, Tomaž. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010,* 17–23

Erjavec, T. (2012) MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1): 131-142.

Evans, N. and S. Levinson (2009) The myth of language universals: Language diversity and its importance for cognitive science. In: *Behavioral and Brain Sciences*, 32(5).

Farrar, S., T. Langendoen (2003) A linguistic ontology for the semantic web. In: *Glot International*, 7(3): 97–100.

Ide, N., N. Calzolari, J. Eckle-Kohler, D. Gibbon, S. Hellmann, K. Lee, J. Nivre, L. Romary (2017) Community Standards for Linguistically-Annotated Resources. In: Ide N., Pustejovsky J. (eds) *Handbook of Linguistic Annotation*. Springer, Dordrecht: 113:165.

Ide, N. and J. Pustejovsky (2010) What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proceedings of the Second International Conference Global Interoperability for Language Resources.* Hong Kong.

Kemps-Snijders, M., M. Windhouwer, P. Wittenburg, S.E. Wright (2009) ISOcat: Remodelling metadata for language resources. In *International Journal of Metadata, Semantics and Ontologies* 8(4): 261–276.

Leech, G., and A. Wilson (1996) *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. EAG--TCWG--MAC/R.* Version of March, 1996. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html [25/03/2018]

Parvizi, A., M. Kohl, M. Gonzàlez, R. Saurí (2016) Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of LREC 2016.*

Petrov, S., D. Das, and R. McDonald (2012) A universal part-of-speech tagset. In *Proceedings of Language Resources and Evaluation Conference 2012*.

Sylak-Glassman, J. (2016) *The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema).* Working Draft v.2. Center for Language and Speech Processing. John Hopkins University. Manuscript. https://unimorph.github.io/doc/unimorph-schema.pdf [23/05/2018]

Zeman, Daniel (2008) Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of Language Resources and Evaluation Conference 2008*.

Acknowledgements

We want to thank Imogen Foxell and Tressy Arts for their thorough considerations of various lexical and grammatical issues around languages that fall quite far from our linguistic knowledge, and which have help shape the ontology put forward here in very positive ways.