Towards Cross-language Application of Dependency Grammar

Timo Järvinen[*], Elisabeth Bertol[*], Septina Larasati[+], Monica-Mihaela Rizea‡, Maria Ruiz Santabalbina○, Milan Souček[*]

[*]Lionbridge Technologies Inc. Tampere, Finland

[+]Charles University in Prague, Czech Republic

‡University of Bucharest, Romania

○University of Valencia, Spain

{timo.jarvinen, milan.soucek}@lionbridge.com, {liz.bertol, septina.larasati, monicamihaelarizea, mrsantabalbina}@gmail.com

Abstract

This paper discusses the adaptation of the Stanford typed dependency model (de Marneffe and Manning 2008), initially designed for English, to the requirements of typologically different languages from the viewpoint of practical parsing. We argue for a framework of functional dependency grammar that is based on the idea of parallelism between syntax and semantics. There is a twofold challenge: (1) specifying the annotation scheme in order to deal with the morphological and syntactic peculiarities of each language and (2) maintaining cross-linguistically consistent annotations to ensure homogenous analysis for similar linguistic phenomena. We applied a number of modifications to the original Stanford scheme in an attempt to capture the language-specific grammatical features present in heterogeneous CoNLL-encoded data sets for German, Dutch, French, Spanish, Brazilian Portuguese, Russian, Polish, Indonesian, and Traditional Chinese. From a multilingual perspective, we discuss features such as subject and object verb complements, comparative phrases, expletives, reduplication, copula elision, clitics and adpositions.

1 Introduction

Dependency-based grammars (DG) have been used in computational linguistics since the formalization of Tesnière's (1959) structural grammar by Hays (1964). The starting point of the work presented in this paper was Stanford typed dependencies (SD) by Marneffe and Manning (2008, revised November 2012). In parallel to our work, the authors of SD have proposed an extended scheme to account for "several linguistically interesting constructions and extend the scheme to provide better coverage of modern web data" (Marneffe & al., 2013), and later, they suggested a revised cross-linguistic typology (Marneffe & al., 2014), and an online discussion forum for Universal Dependencies was opened at http://universaldependencies.github.io/docs/. However, we feel the discussion has not yet fully taken into account the important notions in dependency grammar tradition or the practical requirements of annotation and use of the syntactically annotated data. Our theoretical framework relies on the notions elaborated earlier by Järvinen and Tapanainen (1998).

2 Functional approach for dependencies

The theoretical framework adopted here applies notions inherent in dependency grammar theory to guide the descriptive decisions for particular languages with the aim of producing a universal syntactic annotation scheme that is intuitively clear and that presents the functional syntactic structure in a way that makes it most efficiently available for practical use. A more rigorous framework would help us to address the following (interrelated) deficiencies:

- English bias due to the fact that English was the starting point for the SD.
- Idiosyncracies due to various descriptive traditions as most of the languages under investigation have a long descriptive tradition not related to formal dependency theory.

- Use of notions derived most notably from phrase-structure grammar, though they are not suitable as primitives in DG.
- Pure language-engineering perspective, which may lead to ad-hoc solutions.

The main features of the suggested dependency scheme are:

- The basic syntactic element is a not a word but a nucleus consisting of a semantic head and one or more optional functional words or markers.
- The dependency functions between nuclei are unique within a simple, uncoordinated clause and the inventory of these extranuclear functions is broadly universal.

As elaborated by de Marneffe & al. (2014), SD adopts the lexicalist hypothesis as its first design principle, which regards the word as the fundamental unit in syntax and posits that grammatical relations exist between whole words or lexemes. The authors acknowledge the existence of cases where this assumption fails. First, there are certain types of clitics, which they suggest be treated as independent words even when they are spelled as a single word, following a common practice in many treebanks. Second, there are multi-word lexemes, for which they suggest specific labels such as `mwe`, `name` and `compound` for annotation of the compound parts.

The existence of clitics and multi-word lexemes is not a marginal phenomenon, but it shows that the orthographic word is not suitable as a primitive in DG descriptions. In order to capture what is universal in functional dependency grammar, the notion of nucleus is crucial. It acknowledges the fact that the relations between grammatical markers and content words are different in nature from the relations between content words. The relations within the nuclei are language-specific as there is a large amount of variation in the types of grammatical markers used in different languages. Prototypical markers include adpositions, conjunctions and auxiliaries.

The latest version of SD has adopted a similar view in treating not only auxiliaries but also adpositions as dependents and marking adpositions with a label `case`, which captures the parallelism of adpositional constructions and morphological case.

We discuss the adpositional constructions in detail to illustrate the variation between languages in the choice of adpositional construction versus a specific case marker in the verb complement. In order to achieve a uniform description between languages that takes the functional parallelism fully into account a more thorough revision would be in order. The problem of tokenization is closely related to this issue.

It is a common phenomenon that an orthographic word corresponds to a multiple nucleus; for example, the subject is often incorporated into the verb. Thus, the Spanish token *dámelo* includes three syntactic functions in the verb form: subject, object and indirect object. In practical parsing it may be convenient to use an orthographic word as a primary token, but unless we specify the functional information in the morphological description of the token, the syntactic analysis is not complete.

As both the grammatical markers and syntactic nucleus may consist of several orthographic words, it is convenient to use specific intra-nuclear dependencies linking the parts within them. A common morphological process of reduplication poses problems for the lexicalist hypothesis. The nucleus analysis predicts that there is a continuum from morphological reduplication to full lexicalization.

2.1 Universal dependencies

There are obvious reservations for the universality of the functional dependencies. Presumably, an exhaustive list of functional dependencies may not exist, nor is it necessary to investigate this from the linguistic point of view. As empirical linguists, we only need to list the functions that are applicable to the languages we are analyzing, but we can not assume that all of the universal functional dependencies are applicable to all languages.

From a practical point of view, the most important choices are (i) the selection of the relevant functional categories that need to be covered and (ii) the granularity of the description.

The choice of granularity has an impact both on parsing accuracy and usability of the parsing results. Consider the inventory of adverbial functions as an example. We can use a single functional dependency, adverbial modifier (`advmod`), to annotate optional adverbial modifiers. Alternatively, we could use a more fine-graded set of adverbial functions that includes functions typically distinguished in traditional grammars, such as time, duration, frequency, quantity, manner, location, source, goal, contingency, condition. An obvious advantage of using a large inventory for adverbials is more usable output to various applications requiring even a rough semantic analysis. In fact, a larger set of adverbial roles may improve the parsing accuracy. Though the adverbial modifiers are optional and to a large extent freely combinable with any predicate (save strictly semantic restrictions), it is a commonplace in linguistics that a predicate may have only one non-coordinated adverbial of the same type – a behavior similar to the obligatory arguments or complements. This principle of uniqueness is applicable to practical parsing of adverbials (e.g. to solve the so-called PP-attachment ambiguities) only if all types of adverbial functions, in addition to the complements, are covered in the language model. Recently, Jaworski and Przepiórkowski (2014) have applied a similar idea for assigning approximate semantic roles based on grammatical functions and morphosyntactic features in syntactic-semantic parsing for Polish.

For practical parsing, the uniqueness principle is more important than the distinction of obligatory arguments. An obligatory argument is often missing (being implicit or contextually recoverable), but uniqueness cannot be violated as this would render the clause contradictory or nonsensical. Note that the principle of uniqueness is no longer applicable if several subcategories for unique functional labels are used. For example, the subcategories of subject proposed in SD (`nsubj`, `nsubjpass`, `csubj` and `csubjpass`) are mutually exclusive. As this distinction is automatically recoverable from the linguistic context, it is redundant and it would be advantageous to use only one subject label when doing practical annotation work.

3 Selected linguistic phenomena with reference to SD

3.1 Verb complementation

The grammatical form of complements of verb is governed by the verb. Traditionally, these are considered obligatory versus adjuncts that may occur freely without grammatical restrictions imposed by the verb. From the viewpoint of functional grammar the complements have a specific status. The semantic roles assigned to them are idiosyncratic, depending on the verb. For example, in English, a specific verbs may assign the role of location to a direct object, for example: *They swam a lake.*

The inventory of complement types shows a large amount of language-specific variation, but the core set of complement types is broadly universal. Which complement types are instantiated in a given language can be determined by the uniqueness test. Regarding complement types, our solution was to introduce new dependency relations in our application of the SD model as needed. The cases in point are subject complement (`scomp`) and object complement (`ocomp`), complements that refer to the subject and object, respectively.

Subject complement. The new dependency label `scomp` (subject complement) was introduced to replace `attr`, `cop` and `acomp` (McDonald et al., 2013, p. 3, Table 1; de Marneffe and Manning, 2008), which had been used inconsistently across languages and caused considerable confusion. A subject complement (`scomp`) to a verb has as its antecedent the subject of the clause. In English as well as other languages, it is a widely used grammar term covering the traditional syntactic functions of predicative noun and predicative adjective, frequently, but not exclusively, following a copular verb that links the `scomp` with the subject. `Scomp` occur not only as (pro)nouns (1) and adjectives (2), but also as adverbs (3) as well as prepositional (4) and genitive phrases (5) and in passive structures (6). In languages where `scomp` inflects, adjective `scomp` will agree with the subject in number and gender, as in Romance languages (7).

(1) *¿Qué es esto?*
“What is this?”

(2) *Gold is expensive*.

(3) *Who is there?*

(4) *Sie wurde zur ersten Astronautin Lichtensteins*.
“She became to the first astronaut of Liechtenstein.”

(5) *Sie ist guter Dinge.*
“She is of good things.”

(6) *Il a été nommé president.*
“He has been named president.”

(7) *Quelle est la distance? Jean est petit.*
“Which is the distance? Jean is small.”

Object complement (ocomp) is another dependency label that was introduced to capture complements to the direct object of the verb. It usually occurs in connection with verbs of creating or nominating/naming such as *make, name, elect, paint, call,* etc., which govern at least two complements. The ocomp relation occurs not only with nouns (8) and adjectives (9), but also in prepositional phrases (10). In languages where ocomp inflects, adjective ocomp will agree with the object in number and gender, as in Romance languages (11).

(8) *Te considero una persona inteligente.* (es)
"I consider you an intelligent person."
(9) *We painted the house green.*
(10) *Ich halte die Idee für blöd.* (de)
"I hold the idea for dumb."
(11) *Os críticos acharam o filme fabuloso.*(pt)
"Critics found the movie amazing."

Contrary to scomp, which replaces three previously used labels, ocomp is less a replacement for specific labels than an addition to the dependency relations. Only the previous label acomp (adjective complement) was replaced either by scomp or ocomp, depending on the functional role of the adjective. For example, Tapanainen and Järvinen (1997) include object complement, but de Marneffe and Manning, (2008), do not include anything akin to an ocomp in their list of complements. Prior to the introduction of ocomp, annotators resorted to a variety of solutions, such as acomp if the object complement was an adjective or appos if nominal. Ocomp has been accepted as a viable dependency label by the annotators of all languages in the scope of this project.

Expletive or Topic: The dependency relation expl (expletive) is defined as "a relation that captures an existential *there*". The main verb of the clause is the governor as (12) in de Marneffe and Manning, (2008).

(12) *There is a ghost in the room.*
Expl (is, There)

Also later SD adaptations use this label similarly (McDonald et al, 2013).

Although "expletive" is often defined to include non-referential *it* and equivalents in other languages as in English "*it is raining*" or German "*Es regnet*", by default we adhered to SD guidelines in that expl is used only for equivalents of English existential *there* or non-referential it in clauses or sentences containing a subject in addition to the expletive. Even though there is no semantic subject in structures like *It is raining*, the dummy subject is obligatory in verb-second clauses and it is tagged as nsubj. However in French, we used a broader definition of the notion expletive by making a distinction between the expletive value of the subject and expl as a dependency relation. Therefore, we were able to apply this relation to nouns as well as adverbs or even to prepositions. We needed expl in order to account for a particular dependency relation established by such "empty" words.

For example, we analyzed structures like (12) as expl(a,y) and decided to analyze nsubj(a,il). We also used expl when the subject or direct object position was already filled (for example, in co-referent expressions where we decided that the semantic subject should be analyzed as nsubj (14) expl(est, c'). There were also other situations such as non-negative *ne* (15), euphonic -t: ("*y a-t-il*") (13), to introduce the impersonal subject "*on*" (16), where we had to opt for expl. We have adapted this deprel to the specific situations of French grammar. Our use of expl does not contradict the initial definition. It is only a broader definition, allowing a wider range of uses.

(13) *Il y a un problème.*(fr)
"There is a problem."
(14) *C'est quoi la distance? (*fr)
[Expl]-It-is what the distance.
"What is the distance?"
(15) *Je crains qu'elle ne parte.* (fr)
"I fear she left."
(16) *La situation est bien plus grave que l'on peut imaginer.* (fr)
"The situation is well more serious than one can imagine."

We would like to point out the parallelism between the expletive in subject-prominent languages discussed here and topic in topic-prominent languages like Japanese and Korean, following the distinction by Li and Thompson (1976). From the universal dependency point of view, a single label might be appropriate for both types of languages. The difference is merely the semantically empty topic in subject-prominent languages versus the semantically indeterminate topic in topic-prominent languages.

3.2 Adpositional structures

Typically, adpositional constructions are used as adjuncts. However, in many languages some of the complements are marked with an adposition or a specific case. For example, in English a complement semantically equivalent to an indirect object (`iobj`) is marked with the preposition *to*.

3.3 Comparative constructions

Comparative sentences are those in which a comparison is established. The main clause contains the first term of the comparison, and particular words (like *que* and *como* in Spanish and Portuguese) introduce the second term of the comparison. This second term of the comparison could be a clause or a sentence.

(17) *La empresa realizó trabajos más avanzados que los pioneros de la transmisión.* (es)
"The company accomplished more advanced tasks than the pioneers of the transmission did."

(18) *La guardería no es tan cara como decían.* (es)
nsubj(es, guardería); det(guardería, La); root(es); cop(es,cara); advmod(cara, tan); mark(es,como); advcl(como,decían)
"The nursery school isn't as expensive as they said."

The difference between (17) and (18) is that the first one contains a comparative phrase with no verb in the second term of comparison whereas the latter contains a comparative clause with a verb. This formal distinction has syntactic consequences so the two cases cannot be treated in the same way.

Comparative clauses: Spanish and Portuguese grammars have pointed out that comparative and consecutive clauses are syntactically very similar.

(19) *es tan alto que no cabe por la puerta* (es)
"he's so tall he cannot get through the door"

(20) *era tão alto que batia na porta* (pt)
"he's so tall he cannot get through the door"

Sentences (19) and (20) are formally very close to (18), but the underlying meaning is different. In these cases there is not a comparison, but a cause – consequence relation. This syntactic similarity could be a good reason to consider comparative clauses as `advcl` and, consequently, consider the word that introduces the second term of the comparison as a marker (`mark`).

As shown in the example (18), since the deprel assigned to the clause is `advcl`, the head of comparative clause should be the verb of the main clause, that is, the `root`.

A final observation to be made about comparative clauses is that the preferred POS tag of these markers is CONJ: dictionaries have already pointed this out, and it is consistent with the consecutive – comparative analogy, too.

Comparative phrases: The case of comparative phrases is more complicated because they do not have a verb, and there is thus no parallelism with other kinds of clauses. While it would be possible to analyze these as clauses with omitted verbs, we still would not be able to identify the head.

The most controversial decision was to determine the most appropriate label for the word that introduces the second term of the comparison, because this decision would influence the complete analysis of these phrases.

It was pointed out that *como* could be considered as an adposition (ADP) in some contexts in Portuguese (even if in these cases the dictionaries say it should be a conjunction). In Italian, this marker even selects the oblique case of the pronoun as regular prepositions do, but that is not the case in Portuguese or in Spanish.

(21) *bella come te* (it);
bela como tu (pt);
bella como tu (es)
"beautiful like you"

In Spanish, we can find some examples where the comparative meaning is introduced by an unequivocal ADP:

(22) *es más alto de lo normal*
"he's taller than the average"

Similarly, if we say that *como* is a conjuction functioning as `prep`, the same can be applied to *que* as well:

(23) *mais bela que tu* (pt);
más bella que tu (br)
"more beautiful than you"

Since the final annotation decision was to treat these words as conjunctions with prepositional function, ADP, the complete analysis of the comparative phrase was affected. The corresponding deprel to an ADP should be `prep`, which is always the head of a `pobj`. Consequently, the most appropriate analysis for the comparative phrase is indeed `pobj` and the head would be the verb of the main clause, as in (25).

(24) *La empresa realizó trabajos más avanzados que los pioneros de la transmisión.* (es)
nsubj(realizó, empresa); det(empresa, La); root(realizó); dobj(realizó, trabajos); amod(trabajos, avanzados); advmod(avanzados, más); prep(realizó, que); pobj(que, pioneros); prep(pioneros, de); pobj(de, transmisión); det(transmisión, la)
"The company accomplished more advanced tasks than the pioneers of the transmission did."

Comparative constructions were also discussed by de Marneffe & al. (2013). We agree that their analysis to treat the word that acts as the standard of comparison as the head for the comparative clause or phrase is more adequate from a semantic point of view. This was also the intended analysis in the FDG description (Järvinen & Tapanainen 1997):

(25) *There are monkeys more intelligent than Herbert.*
modifier(more,than); pobj(than,Herbert)

This analysis is further corroborated by typological evidence. For example, in Korean the comparative particle 'more than' is a single unit that attaches to the object of comparison (Yeon & Brown, 2011):

(26) 러시아가 한국보다 더 크다.
Russia-TOPIC Korea-THAN big
"Russia is bigger than Korea."

3.4 Clitic particles

New POS tag	Description
VERBPRONACC	verb + accusative clitic
VERBPRONDAT	verb + dative clitic
VERBPRONDATACC	verb + dative clitic + accusative clitic
VERBPRT	verb + verbal morpheme (PRT)
VERBPRTPRONACC	verb + PRT + accusative clitic
AUXPRONACC	auxiliary verb + accusative clitic
AUXVPRT	auxiliary verb + PRT

Table 1. List of new POS tags created for Spanish.

Even in closely related languages such as Portuguese and Spanish, which exhibit a broadly similar behavior of clitics, the differences in orthography make the practical analysis for the latter more challenging. In Spanish, the enclitic pronouns are orthographically attached directly to the verb form and consequently, a mechanical tokenization of the complex word form is not possible as in Portuguese, which uses a hyphen in this context. Rather than attempting to tokenize the Spanish clitics separately, we used an extended set of POS labels for Spanish as illustrated in Table 1, so that there would be no loss of information as compared to the analysis of other Romance languages. This descriptive solution is made for convenience, but note that the functional description is not compromised. It is a purely technical question whether to use a single POS label or a main POS label with separate morpho-syntactic descriptors to encode the values for incorporated syntactic functions. A more complete syntactic description for the example *dámelo* would be VERB + Subj_Sg2 + Dat + Acc, thereby making the information available for conversion to a proper functional DG description showing the three nuclei as direct dependents of the verbal nucleus.

3.5 Multi-word expressions

As for mwe modifiers, we have consistently annotated idiomatic word combinations whose internal structure is not relevant for the functional analysis by using the other existing dependency relations and POS (regent–subordinate) combinations that were permitted for each language.

In de Marneffe and Manning (2008), the mwe dependency relation implies a closed set of items (restricted mainly to function words). By convention, the internal head of the mwe relation is consistently analyzed, across languages, as the rightmost element of the structure.

We kept a list of possible mwe candidates that were approved during the project for all languages. For some Romance languages (e.g. French), it was convenient to define patterns of mwe, as opposed to a plain list of these. Generally, idiomatic combinations that consisted of preposition + (preposition) + noun, pronoun, adjective, adverb or infinitive were analyzed as surface prep and pobj or/and pcomp structures; mwe was used for semantically opaque expressions that mostly included structures consisting of adverb, noun or conjunction + adposition or conjunction, for example in Spanish *mientras que* mark(*,que), mwe(que, mientras) POS: CONJ, CONJ; *para que* mark(*, que), mwe(que, para) POS: ADP, CONJ; in Brazilian Portuguese *até que* mark(*, que); mwe(que, até) POS: ADP, CONJ; and French *avant/afin de,* see (30); *pour que* mark(*,que), mwe(qu',pour) POS:ADP, CONJ.

Additionally, in deciding whether a multi-word structure is analyzable, we also had to consider the relation that needed to be established between the components of the structure and the external elements. For example, some French ‘locutions prépositives’ of the type preposition + noun that are followed by a nominal are analyzed as mwe since there is no acceptable interpretation for the following nominal in case we analyze the prepositional structure as prep and pobj:

(27) *Ils sont tous venus, à part Christian.*
prep (venus, part); mwe(part, à);
pobj(part,Christian).
“They are all come, except Christian.”

(28) *Cet objectif peut être réalisé à travers les règles à fixer par la Commission.*
prep(réalisé,travers); mwe(travers, à);
pobj(travers, règles).
“This objective can be realized by means of the rules to fix by the commission.”

It can be noticed that the governor of a multi-word expression annotated as mwe takes the head of the expression as a subordinate, using a dependency relation which describes the relation between the governor and the mwe. Examples from French:

(29) *en tant que*
prep(*,tant), mwe(tant,en), mwe(tant,que)
POS: ADP/ADV/CONJ
“as”

(30) *avant de*
mark(*,de), mwe(de,avant)
POS:ADV/ADP
“before”

(31) *beaucoup de*
det(*,de), mwe(de,beaucoup);
POS:ADV/ADP
“a lot of”

In (31) the pattern comprises of *beaucoup, plein, bien, peu, tant, assez, plus, advantage* and *sufficamment.*

Sometimes, mwe might imply a head which is morphologically different from the function of the whole structure. For example, the French mwe *peut-être* has an adverbial value. This implies that the head of the mwe, *être*, which is actually a verb, becomes subordinated by an advmod deprel to the governor of the multi-word the structure:

(32) *Criton sait que Socrate est aussi fidèle que lui et il pense que si Socrate ne se sauve pas pour lui - même , peut - être se sauvera - t - il pour ses amis.*
advmod(sauvera, être); mwe(être, peut).
“Criton knows that Socrates is as faithful as him and he thinks that if Socrates not himself saves not for himself, maybe himself will save he for his friends.”

Spanish and Brazilian Portuguese also permit a noun (which was the head of a mwe functioning as a conjunction) as a subordinate in a cc deprel:

(33) *Los objetivos de los aliados , sin embargo , diferían.* (es)
cc(diferían, embargo); mwe(embargo,sin).
“However, the aims of the allies differed.”

Similarly with a verb:

(34) *Es decir, un jugador puede jugar como un WHM*. (es)
cc(jugar, decir); mwe(decir, es).
“That is, a player can play as a WHM.”

(35) *Denomina - se oblíquo quando não é um cone reto , ou seja , quando o eixo é oblíquo ao plano da base.* (pt)
cc(é, seja); mwe(seja, ou).
“A cone is called oblique when it's not upright, that is, when its axis is oblique to the plane of its base.”

3.6 Elision

Dependency theory is inherently verb-centered. Therefore elision of a verb poses a descriptive problem that could be solved either by (a) inserting an empty node (represented as EMP in the example below), which assumes the functions of the elided element or (b) raising an existing element to the position of the elided node. The examples for solution (a) and (b) are provided in (36) and (38), respectively, for comparison.

(36) *Beliau seorang penerbit.*
PRON DET NOUN
root(*, EMP); nsubj(EMP, beliau); dobj(EMP, penerbit); det(penerbit, seorang)
“He is a publisher.”

The former solution (a) is not plausible if the purpose is to provide a surface-syntactic functional description rather than an abstract deep-syntactic representation of an elliptic sentence. Positing an abstract representation by analogy, as ellipsis is often described in traditional grammar, is questionable as a syntactic analysis in the sentence level and computationally more challenging as it would mean that the parser should somehow be able to map the non-elliptic construction to the elliptic construction to produce the intended analysis. Therefore, achieving the best possible analysis between the actual elements in the sentence or sentence fragment is strongly preferred.

Elision of a copula in present tense is standard in Russian and it may appear in informal registers (speech transliterations) in Indonesian.

Our examples are from Indonesian, which uses copulas to link a subject to nouns, adjectives, or other constituents in a sentence. There are three copula constructions found in our data. These constructions are sentences with a copula, sentences with a dropped copula, and sentences with a verb that acts like a copula. For some of these constructions we use the `scomp` deprel to create the link between the constituents. These copulas are not auxiliary verbs, hence they are not annotated as AUX, but instead they are annotated as VERB.

Sentences with copulas: There are two copulas in Indonesian, *adalah* and *ialah*. They have the same function and can be used interchangeably. These copulas cannot be negated. We use the `scomp` deprel to link the subject and the other constituents that surround the copula.

(37) *Beliau adalah seorangpenerbit.*
PRON VERB DET NOUN
root(ROOT, adalah); nsubj(adalah, Beliau); scomp(adalah, penerbit); det(penerbit, seorang)
“He is a publisher.”

Sentences with dropped copulas: In some cases, especially in spoken Indonesian, the copulas can be dropped. The sentence can be negated.

(38) *Beliau seorang penerbit.*
PRON DET NOUN
root(*, Beliau); scomp(Beliau, penerbit); det(penerbit, seorang)
“He is a publisher.”

Sentence with copula-like verb: The verb that acts like a copula is the word *merupakan*, which links the subject to the other constituents. This verb can be negated. The sentence is annotated as a usual Subject-Verb-Object (SVO) structure in Indonesian without the `scomp` deprel.

(39) *Beliau merupakan seorang penerbit.*
PRON VERB DET NOUN
root(*, merupakan); nsubj(merupakan, Beliau); dobj(merupakan, penerbit); det(penerbit, seorang)
“He is a publisher.”

3.7 Reduplication

Another common morphological process that is of interest here is reduplication. This structure is found in our data in Indonesian and traditional Chinese (Larasati, 2012, Wang, 2010). Reduplicated forms were tokenized into separated tokens.

To accommodate this phenomenon, a new dependency relation, `redup`, was introduced to link the reduplicated token. Depending on the language, a reduplicant may copy either from the right or from the left and the governing head is either to the left or to the right, respectively. We used the leftmost token as the head (Wang, 2012).

One of the uses of reduplication in Indonesian is to indicate plurality, e.g. the word *senapan-senapan* (n. 'riffles', lit. riffle-riffle). Some reduplicated nouns are lexicalized, e.g. *langit-langit* ('ceiling; palate' < langit, 'sky').

From the functional point of view, `redup` is an intranuclear link. The analysis may not distinguish fully between lexicalized and non-lexicalized instances, though in the former case a single-token analysis would be more appropriate.

For Traditional Chinese, one of the uses of reduplication is to intensify the degree to which the property denoted by the adjective holds, e.g. the word "小小"(adj. very small, lit. small small). In the data, the word is tokenized into two tokens "小"and "小".

4 Conclusion

Applying a strict linguistic theory would assist linguists in choosing between alternative annotations more consistently and efficiently.

It is not possible to achieve a consistent and descriptively adequate cross-lingual description without a consistent theoretical framework. A plain eclecticism would only lead to a proliferation of the grammatical descriptors.

Functional syntactic descriptions have gained ground in computational applications. The notions of phrase-structure grammar are tied to the form of a particular language, and as there is a need to cover more and more new languages of various types, functional descriptions that capture the implicit semantic parallelisms between languages provide an even more adequate framework for practical work and practical applications.

Acknowledgements

We wish to thank the three anonymous reviewers for their comments on the submitted version. The data annotation was conducted in a Multilingual Data Annotation Project executed for Google. This research was partially supported by SVV project No. 260 224 of the Charles University in Prague.

References

Bernard Comrie. 1989. *Language universals and linguistic typology*. 2nd edition. Chicago: University of Chicago.

David G. Hays. 1964. Dependency theory: A formalism and some observations. Language, 40:511-525.

Charles N. Li and Sandra A. Thompson . 1976. *Subject and Topic: A New Typology of Language"*. In Charles N. Li. *Subject and Topic*. New York: Academic Press.

Wojziech Jaworski and Adam Przepiórkowski: Syntactic Approximation of Semantic Roles. 2014. In *Proceedings of 9th International Conference on NLP, PolTAL 2014*. Pp. 191-201.

Timo Järvinen and Pasi Tapanainen. 1997. A *Dependency Parser for English*. Technical Reports, No. TR-1. University of Helsinki.

Timo Järvinen and Pasi Tapanainen. 1998. Towards an Implementable Dependency Grammar. In: *Proceedings of Dependency-Based Grammars,* (eds.) Sylvain Kahane and Alain Polguère, Université de Montréal, Quebec, Canada.

Septina Dian Larasati. 2012. IDENTIC Corpus: Morphologically Enriched Indonesian-English Parallel Corpus. In *Proceedings of LREC 2012*, page 902-906.

Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Samuel R. Bowman, Timothy Dozat and Christopher D. Manning. 2013. More constructions, more genres: Extending Stanford Dependencies. In: *Proceedings of the Second International Conference on Dependency Linguistics* (Depling 2013).

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Gintner, Joakim Nivre and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In: *Proceedings of LREC 2014.*

Marie-Catherine de Marneffe, and Christopher D. Manning. 2008 (revised Nov. 2012). *Stanford typed dependencies manual.*

Ryan McDonald, Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu, Castelló, N., Lee, J.2013. Universal Dependency Annotation for Multilingual Parsing. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Katarzyna Marszałek-Kowalewska, Anna Zaretskaya and Milan Souček. 2014. Stanford Typed Dependencies: Slavic Languages Application. In *Proceedings of 9th International Conference on NLP, PolTAL 2014*. Pp. 151-163.

Milan Souček, Timo Järvinen and Adam LaMontagne. 2013. Managing a Multilingual Treebank Project. In: Proceedings of the Second International Conference on Dependency Linguistics. (Depling 2013).

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Librairie C. Klincksieck, Paris.

Jaehoon Yeon and Lucien Brown. 2011. Korean: A Comprehensive Grammar. Routledge.

Wang, Zhijun, 2010. The Head of the Chinese Adjectives and ABB Reduplication. NACCL.