Bringing Information Structure to Universal Dependencies

Nikolett Mus[1], **Andrew Dyer**[2], **Claudia Corbetta**[3], **Sylvain Kahane**[4]
[1] ELTE Research Centre for Linguistics, [2] Saarland University,
[3] University of Bergamo/Pavia, [4] Paris Nanterre University
mus.nikolett@nytud.elte.hu, andrew.dyer@uni-saarland.de,
claudia.corbetta@unibg.it, skahane@parisnanterre.fr

Abstract

The aim of the paper is to present a first attempt at annotating Information Structure roles in syntactic treebanks of the Universal Dependencies collection, discussing theoretical considerations and practical methodological questions while presenting our core annotation principles. We focus on constructions in which Topic or Focus is overtly marked through (morpho)syntactic means. The proposed annotation is illustrated using examples from five languages: Wolof, Japanese, Tundra Nenets, Hungarian, and Italian.

Keywords: information packaging, Information Structure, Topic, Focus, Contrast

1. Introduction

The aim of this study is to present a recent initiative and ongoing work on the annotation of Information Structure roles in existing and future Universal Dependencies (UD) treebanks (De Marneffe et al., 2021). As discussed in greater detail in Section 3, we understand Information Structure as the organisation of utterances relative to the discourse context. Our work focuses on annotating the roles that individual sentence elements play in Information Structure, rather than attempting to represent the full Information Structure of the sentence.

In line with Topic–Focus Articulation (henceforth TFA) (Hajičová et al., 1998; Firbas, 1992), we adopt Topic as the element that identifies what the sentence is about and anchors the utterance in the shared discourse. We depart from TFA, however, in our treatment of Focus. Rather than annotating the entire focus domain, we mark only a single element that can be interpreted as Focus, that is, the element that is emphatic or carries the most prominent informational or semantic weight within the utterance, if such an element is present. This prominence is primarily understood in communicative terms, though it may also be reflected prosodically or structurally. In this sense, our notion of Focus aligns with approaches that treat focus as a distinguished element associated with alternatives or contrast (Rooth, 1992; É. Kiss, 1998; Krifka, 2008). In addition to Topic and Focus, we introduce a third notion, Contrast, capturing elements interpreted against a set of alternatives and highlighting a distinction or opposition within the discourse.

A substantial body of cross-linguistic and typological research has examined how languages encode discourse-related distinctions. These studies document considerable variation and emphasise the diversity of strategies employed across languages, ranging from prosodic marking to morphological and syntactic means (Neeleman et al., 2009; Zimmermann and Féry, 2010). In some cases, languages develop specialised constructions whose primary function is the marking of Information Structure roles. In others, elements or constructions with independent grammatical functions are extended to serve Information Structure-related purposes as well. Together, these patterns demonstrate that Information Structure can systematically shape surface syntactic and morphological structure.

Despite this, current UD treebanks, while already invaluable resources for syntactic typology and comparative research, remain limited in that they represent syntactic structures largely in isolation from discourse-level information. As a consequence, the influence of Information Structure on word order, constituent realisation, and constructional variation often remains implicit and inaccessible within existing annotation frameworks. This limitation hinders both fine-grained typological comparison and the interpretation of certain syntactic patterns, particularly in languages whose syntax and grammar have not yet been exhaustively described or analysed.

This is not to say that there have been no attempts to annotate Information Structure in existing corpora, or to create IS-specific corpora (these efforts are discussed in more detail in Section 2). Our approach, however, departs fundamentally from these earlier initiatives. Rather than attempting to identify and annotate complete Information Structural configurations, we restrict our annotation to explicitly marked elements, that is, linguistic units that directly signal, or indirectly indicate, an Information Structure role.

The primary goal of our work is to define Information Structure roles in a precise, operational manner. While this definition is necessarily pre-

liminary and may be refined based on the data, it provides a foundation for both theoretical and practical annotation guidelines. Introducing an Information Structure layer into treebanks can enhance the descriptive and explanatory power of syntactic annotation, enable systematic cross-linguistic comparison of discourse-sensitive structures, and support the refinement of existing typological generalisations. At the same time, explicit annotation of discourse-marked constructions allows these elements to be filtered out, providing a clearer view of a language's unmarked or basic word order patterns.

Beyond its relevance for theoretical linguistics, this effort also has practical and computational applications. Explicit annotation of Information Structure roles can inform research in translation studies and machine translation, support the development of language learning tools, and be particularly valuable for endangered or under-resourced languages, where annotated corpora can contribute to documentation and revitalisation efforts.

As this work is at an early stage, the present paper focuses on outlining our conceptual foundations, discussing the theoretical and practical principles, and illustrating the proposed approach through a small number of case studies. These examples demonstrate the methodological considerations involved in deciding what to annotate and how. Thus, this paper emphasises conceptual clarification and annotation design rather than large-scale annotated data.

2. Related Work

Several previous initiatives have sought to incorporate Information Structure or related discourse-level distinctions into corpus annotation frameworks. These approaches differ considerably in their theoretical assumptions, the units they annotate, and the way Information Structure is represented in relation to syntactic structure.

A number of proposals aim to provide a comprehensive annotation of Information Structure, typically operating over constituents rather than individual tokens. Dipper et al. (2007), for instance, develop a framework for annotating both Information Structure and information status across languages, introducing coarse- and fine-grained labels (e.g. *aboutness* vs. *frame-setting* topics, *new information* vs. *contrastive* focus), along with pragmatic diagnostics for their identification. Similarly, Mírovský et al. (2013) extend the Prague Dependency Treebank (Bejček et al., 2013) with Topic–Focus annotation grounded in the Prague School tradition Hajičová et al. (1998), and Bohnet et al. (2013) introduce a comparable Topic–Focus annotation layer for the Penn Treebank.

Other work addresses more specific aspects of Information Structure or its formal and computational modelling, rather than proposing full annotation schemes. Song (2017) provide a multilingual HPSG-based formalisation of Topic–Focus Articulation, with a detailed taxonomy of Information Structure types and their marking strategies, and demonstrate its usefulness in machine translation. Stede and Mamprin (2016) introduce an Information Structure annotation layer in the Potsdam Commentary Corpus, focusing in particular on fine-grained distinctions among types of Topic. Booth (2022) take a methodological perspective, outlining desiderata for Information Structure annotation, especially with respect to discourse-sensitive phenomena in subordinate clauses, and illustrate their proposal on a corpus of Historical Low German.

Some of these resources have been integrated into broader infrastructures such as CorefUD (Novák et al., 2025), which brings together coreference corpora in the CoNLL-U format and enables the joint analysis of coreference, Information Structure, and dependency syntax.

While these approaches provide valuable insights into Information Structure annotation, they are often challenging to reconcile with token-based frameworks such as Universal Dependencies, which do not explicitly represent constituent structure. By aiming to annotate entire IS configurations, they can introduce ambiguity and require discourse-specific interpretation. In contrast, our role-based approach focuses solely on elements that are overtly marked, treating them as anchor points for IS roles. This strategy avoids speculative assignment, limits potential misinterpretation, and maintains compatibility with UD while supporting cross-linguistic comparison.

3. The concept and definition of Information Structure roles

In this section, we outline the conceptual foundations of Information Structure as it is treated in our annotation framework.

Information Structure concerns how utterances are organised relative to the discourse context, in particular how speakers indicate what a sentence is about and what information is asserted regarding that element. Theoretically, sentences can be conceptually divided into two functional fields: one that establishes the discourse anchor, signalling what the utterance is about, and one that carries the assertive contribution, specifying the new information associated with that anchor. While this division provides a useful generalisation, there are non-prototypical constructions in which a clear discourse anchor is absent or difficult to identify (e.g., in existential clauses such as *There are ghosts*).

A variety of theoretical frameworks have been

proposed to capture this two-part division of sentences into a discourse anchor and an assertive contribution, including Topic–Comment (Gundel, 1988), Given–New (Chafe, 1976), Theme–Rheme (Weil, 1879; Firbas, 1964; Mel'čuk, 2001), and Common Ground-based approaches (Stalnaker, 2002). Each of these frameworks offers valuable insights into how speakers organise utterances relative to discourse, yet they differ in their conceptualisations of key notions, their units of analysis, and their formal assumptions. This diversity, while theoretically rich, presents practical challenges for annotation, particularly within the constraints of a Universal Dependencies framework, where abstract fields or domains may not align straightforwardly with syntactic structures representing dependency relations between syntactic units.

Due to the diversity of conceptual frameworks and the fact that definitions are often either underspecified or fail to adequately account for exceptional and ambiguous cases, previous attempts to annotate Information Structure have not yet fully resolved these issues (Paggio, 2006; Dipper et al., 2007; Ritz et al., 2008; Lüdeling et al., 2016). While these approaches have provided important insights, their theoretical assumptions and representational choices have limited their applicability in a consistent annotation setting.

Importantly, our goal is not to construct a dedicated Information Structure corpus by selectively collecting prototypical constructions. Rather, we aim to annotate Information Structure roles within existing corpora, marking those instances where relevant roles can be identified with sufficient clarity. Therefore, we adopt a hybrid approach, informed by insights from the existing literature, but tailored to the constraints and goals of dependency-based annotation. Within this framework, we define the core Information Structure roles as follows:

(1) **Topic**: the element that identifies what the sentence is about and anchors the utterance in the shared discourse.[1].

(2) **Focus**: the most prominent element in an utterance, i.e., the element that is made emphatic or highlighted.[2]

[1]This definition follows the Topic notion in TFA as proposed by Hajičová et al., 1998; Firbas, 1992.

[2]Here, emphasis is understood in communicative terms rather than purely prosodic or structural, although prosody or structure may also realise it. Unlike the classical TFA approach, we do not annotate the entire focus domain, but rather the single element within it, if such an element is present. This perspective aligns with approaches that treat focus as a marked element associated with alternatives or contrastive relations (Rooth, 1992; É. Kiss, 1998; Krifka, 2008).

At the current stage of analysis, we do not further subdivide these categories into more fine-grained types. Although this simplification may obscure certain distinctions, it provides a necessary starting point for developing a consistent and operational annotation scheme.

In addition to Topic and Focus, we introduce Contrast as a further dimension of annotation. Contrast is not treated as an independent Information Structure category, but as a property that can be associated with either Topic or Focus, reflecting the presence of alternatives in the discourse.

We do not assume that every sentence contains a Topic or Focus, and we annotate only elements that are explicitly marked. By restricting annotation to these identifiable elements rather than representing full Information Structure configurations, we obtain anchor points that can be systematically related to syntactic structure. This approach allows us to examine how syntactic positions correlate with discourse functions and to empirically evaluate assumptions that have often been proposed without extensive corpus-based evidence. At the same time, it provides a minimal, reproducible model with clearly defined roles that can be consistently applied across languages, enabling both the reliable identification of Information Structure roles and the exploration of cross-linguistic patterns of information packaging. In this way, the framework lays the groundwork for empirically grounded typological generalisations concerning Topic, Focus, and Contrast.

4. Annotation of Information Structure roles

In our annotation scheme, both the units that carry Information Structure roles and the morphological or syntactic elements that mark these roles are annotated. This captures the overt realisation of the role, allowing the annotation to represent both the discourse function and its grammatical encoding.

In some cases, these two levels coincide, for example, for inherently marked IS roles or when bound morphemes are part of the constituent itself and directly indicate its IS function, such as focus particles. In other cases, an entire string of tokens (a phrase) may collectively realise a role. In such instances, we annotate the governor (head) as an anchor for the corresponding role, which can then be related to syntactic structure or combined to approximate the full phrasal unit.

As the UD framework is inherently token-based, we use the `MISC` column to record Information Structure roles ensuring that discourse-functional distinctions are recorded without conflicting with the syntactic and morphological core of UD.

For the annotation of role-bearing constituents, a

single umbrella label, `ISRole`, is used. The values indicate the discourse role: `ISRole=Top` for Topic, `ISRole=CTop` for contrastive Topic, `ISRole=Foc` for Focus, and `ISRole=CFoc` for contrastive Focus. This follows UD's feature/value design: the label functions as the category (analogous to `Animacy`, `Number`, or `Case`), and the value indicates the specific role.

Elements that indicate an IS role are annotated with the labels `ISMarker[(C)Top]` and `ISMarker[(C)Foc]`.

Morphological and syntactic markers of IS roles are annotated in accordance with UD principles, with attention to whether the specific grammatical function or construction type is identifiable.

For morphology, when the grammatical function of a (bound) morpheme is known, it is recorded explicitly. For example, an accusative case marker functioning to signal a topicalised constituent can be annotated as `ISMarker[Top]=Case`. In instances where only the morphological type is known, but the precise grammatical function cannot be determined, we assign an underspecified value, such as `ISMarker[Top]=Morph` or `ISMarker[Foc]=Morph`, thereby preserving the morphological information without overcommitting to a specific function. For clarity and to retain more detailed information, we recommend optionally providing a sublabel that records the surface form of the morph, e.g., `ISMarker[Foc]Form=`⟨`form`⟩.

For syntactic marking, we distinguish between construction-based and position-based strategies. Construction-based marking is applied when a particular syntactic construction signals an IS role. When the construction type is identifiable, the root of the construction is annotated with a specific value, for instance, `ISMarker[Top]=Inversion` or `ISMarker[Foc]=Cleft`. When the precise construction is unknown or cannot be determined with confidence, an underspecified value, `ISMarker[Top]=Construction` or `ISMarker[Foc]=Construction`, is used. Position-based marking is applied when a constituent with an IS role occupies a dedicated syntactic position, such as preverbal focus in Hungarian; in this case, the constituent is annotated with the value `ISMarker[Foc]=Position`.

This approach allows us to capture overt morphological and syntactic cues of Information Structure systematically, while providing flexibility to accommodate uncertain or partially analysed phenomena, ensuring consistency and cross-linguistic applicability in the annotation process.

5. Principles of Information Structure annotation

On the basis of our definitions and concepts, we formulated a concise set of principles to guide the annotation of Information Structure roles ensuring consistency, cross-linguistic comparability, and compatibility with the Universal Dependencies framework. First, only explicitly marked elements are annotated, covering overt morphosyntactic and structural cases, while unmarked elements are left unannotated. Second, annotation is function-oriented: structurally distinct constructions receive the same label if they fulfil the same role. Third, comparability across languages is achieved by treating universality at the level of functional roles rather than formal realisation, supporting typological exploration and cross-linguistic analysis. Fourth, contextual information is considered to confirm whether an element is overtly marked as Topic, Focus (or Contrast), without inferring roles in the absence of clear evidence. Fifth, annotation is limited to the core roles, i.e. Topic, Focus, and Contrast, while finer subtypes, such as frame-setting versus aboutness topics (Fery and Krifka, 2008), are not distinguished. Sixth, both the units carrying Information Structure roles and any morphological or syntactic markers signalling these roles are annotated, capturing the discourse function alongside its grammatical encoding. Seventh, when an entire phrase realises an Information Structure role, only the head of the phrase is annotated. Finally, all annotations are encoded in the `MISC` column, which is designed for extra, non-syntactic functional information in UD.

6. Case studies

In this section we provide case studies of some languages in Universal Dependencies (or pending release) with known morphosyntactic mechanisms of Information Structure. We do not consider this selection of languages to be fully representative of all possible strategies for marking Information Structure. Instead, the sample was chosen to capture a diversity of marking practices, illustrating how different languages employ syntactic and/or morphological strategies to encode Information Structure roles. By including a range of strategies here, the examples allow us to demonstrate annotation in practice. We note that, for some of the languages included in this study, the available UD treebanks are still relatively small. Where necessary, the examples presented here are drawn from our own data. Comparable instances can often be found in the treebanks, but minimal pairs or carefully selected examples are chosen to ensure clarity and facilitate explanation.

6.1. Wolof (Niger-Congo)

Wolof is a language where focalisation has been grammaticalised (Robert, 1991, 2000) and almost all sentences contain a particle indicating the focus of the sentence. Only 12% of the sentences in UD_Wolof-WTB treebank (Dione, 2019) have a main clause without a particle (Bondéelle and Kahane, 2021) and this constrained construction is mainly used in narrative sequences (Robert, 1991). The focusing particles of Wolof are verbal particles, which are analysed as complementizers by Torrence (2005) and annotated AUX in the UD_Wolof treebank developed by Dione (2019). Wolof has a very rigid word order, but the word order depends on the particle that has been selected. Two particles, *a* and *la*, are respectively used to focus the subject (3a) and a complement (3b); they occupy the second position, the first position being occupied by the focused element. Dislocated units or modifiers can precede the first position, but are prosodically detached in spoken data and generally separated from the sentence nucleus by a comma in written data.

(3) a. *Bu jëkk, nguur*
when be_first, power
ISRole=CTop
ISMarker[CTop]=Dislocation
googu, ay socé ak ay
CL.ANAPH IND.CL Mandinka with IND.CL
ISRole=Foc
séeréer, ñu =a
Serer, S3PL =PART
ISMarker[Foc]=Morph
ko yor =oon.
O3SG hold =PAST.

'Originally, this power, the Mandinka and Serer, it was them that held it.' (WTB-3)

b. *Noonu it la*
CL.ANAPH too PART.3SG
ISRole=Foc ISMarker[Foc]=Morph
àtte ci am tudd.
decide LOC CL.IND name

'That's also the way he decides on a name.' (WTB-1902)

The realisation of the subject is very constrained and it is generally dislocated as in (3a) (Bondéelle and Kahane, 2020). The dislocation of other elements marks a topicalisation.

Interestingly, there are also two particles, *na* and *da*, that respectively focus the verb and the whole sentence. The particle *na* is not recognised as a focus particle by most authors, but Bondéelle and Kahane (2021) postulate that it could be a translation effect, because the focus of the verb is not marked and is the default interpretation of a sentence in languages used for the description of Wolof, such as French and English. The particle *na* occupies the second position, the verb being in the first position, while *da* occupies the first position. The contrast between *na* and *da* can be illustrated by the following examples from Robert (2000).

(4) a. *Ragal na ko.*
be_afraid PART him.

'He is afraid of him.', lit. it's the fear he feels towards him.

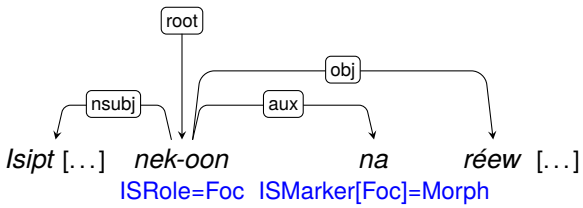
b. *Dafa ragal.*
PART.3SG be_afraid.

'He is a coward.'

The particle *na* is used to focus the imperfective auxiliary *di*, giving the locution *dina* that marks the future tense. The following example shows a focus on the past tense.

(5) *Isipt gu yàgg ga*
Egypt CL.REL last(V) CL.DEF
nekk-oon na réew mu
be-PAST PART.S3SG country CL.REL
woomle, […]
be_prosperous, […]

'Ancient Egypt WAS a prosperous country, […]' (WTB-245)



The particle *da* can be analysed as a sentence-focus, in the sense of Lambrecht and Polinsky (1997). When it is used with an action verb, it generally appears in a sequence of sentences and takes a causal value, as in (6) and (7a) (Robert, 2000).

(6) *Sama jëkkër nekk-u fi,*
my husband be-NEG.S3SG here,
dafa dem àll ba.
PART.S3SG go forest CL.DEF

'My husband isn't here; he's gone into the bush.'

The negation has different realisations according to the focus. It can be realised by the suffix *-u*, the negative value of *na* (7a); by the particle *du*, the

negative value of *da* (7b);[3] or by a negation *-ul* on the verb when the focus is elsewhere (7c).

(7) a. *Tóx-u-ma, da-ma-y fo.*
smok-NEG-1SG, PART-1SG-IMPF play.
'I'm not smoking, I'm playing (with the cigarette).', lit. what I'm doing isn't smoking, it's playing.

b. *Du-ma tóx.*
NEG.PART-1SG smoke.
'I don't smoke, I'm not a smoker.'

c. *ma =a tóx-ul.*
1SG =PART smok-NEG.
'It's me who don't smoke.'

UD_Wolof-WTB already has a feature `FocusType`, though this is unique to the language. The values are `Subj` for *a*, `Comp` for *la*, and `Verb` for *da*. We propose to associate the `Verb` to *na* and to put `Sent` on *da*. This feature is partly redundant with our annotation of `ISRole`, but not completely, because `ISRole[foc]` will be on the verb with both *na* and *da* and we do not specify whether it applies on the verb only or a larger domain.

6.2. Tundra Nenets (Uralic, Samoyedic)

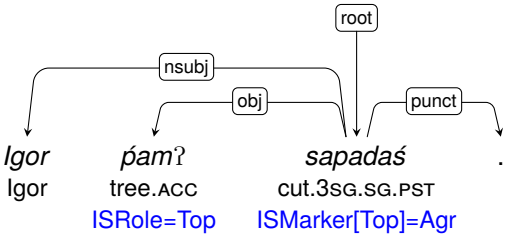
Tundra Nenets, an endangered, and relatively underdescribed language spoken in the Arctic regions of Russia, appears to employ morphological strategies for encoding Information Structure roles. On the basis of existing descriptions, both Topic and Focus – or at least certain subtypes thereof – can be expressed through morphological marking (Nikolaeva, 2014).

Accordingly, one particular type of Topic – specifically, the lexical object of the clause when it serves a topical function – appears to be morphologically marked in the language through the so-called objective verb conjugation expressed by a suffix added to the verb. This marker simultaneously encodes agreement with both the subject and the object. The objective conjugation system contrasts with forms that encode agreement with the subject alone, and it is typically restricted to third-person objects (Dalrymple and Nikolaeva, 2011; Nikolaeva, 2014). The minimal pair presented below illustrates this contrast: in (8a) the transitive verb agrees only with its subject in person and number, and in (8b) it agrees both with its subject and with the third-person topical object in number.

(8) a. *Igor ṕamʔ sapaś*
Igor tree.ACC cut.3SG.PST
'Igor cut a/the tree.'

b. *Igor ṕamʔ sapa-da-ś*
Igor tree-ACC cut-3SG.SG-PST
'(As for a/the tree) Igor cut a/the tree.'

In such cases – that is, when the verb agrees with an object bearing a Topic role – the object receives the label and value `ISRole=Top`. Since, within the UD framework, agreement morphology is not segmented as a separate morpheme, the verb is annotated with the label `ISMarker[Top]=Agr`. The example above can therefore be represented by the following UD tree.



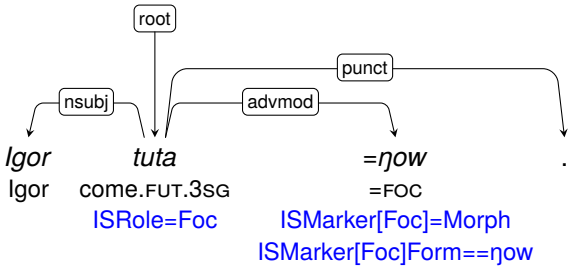
In addition to the marking of object Topics, Tundra Nenets also exhibits clitics and/or affixes that serve to encode Focus, or at least certain subtypes of foci (Nikolaeva, 2014). One such element is the so-called focus marker *=(ŋo)w* / *=ej*, which attaches to verbs and appears as the final morpheme in the word form. This affix functions to emphasise the action expressed by the verb and, more specifically, to highlight the truth value of the proposition, see (9) and Nikolaeva (2014).

(9) *Igor tu-ta=ŋow*
Igor come-FUT.3SG=FOC
'(Indeed) Igor WILL come.'

In this instance, the verb bearing the suffix is annotated as `ISRole=Foc`, while the overt marker itself is labeled `ISMarker[Foc]=Morph`. As mentioned above, for clarity and to retain detailed information, the surface form of the marker can optionally be recorded using the sublabel `ISMarker[Foc]Form==ŋow`[4]. This is illustrated in the following UD tree:

[3] *da* and *du* are generally analysed as constructed on the imperfective auxiliary *di*, *da* as *di* plus the focus marker *a* and *du* as *di* plus the negative focus marker *-u* (Sauvageot, 1965).

[4] Given its position as a closing suffix on the verb and its role in marking verum (emphasis on the truth of the proposition), we annotate the clitic as an adverb (ADV) and assign it an `advmod` relation to the predicate.



(12) a. *Niku wa tabenai*
meat TOP eat-NEG

'I don't eat meat.' (lit: 'As for meat, [I] don't eat [it].')

b. *Niku o tabenai*
meat ACC eat-NEG

'I don't eat meat.' (lit: '[I] don't eat meat.')

6.3. Japanese (Japonic)

In Japanese, phrases within a sentence are typically marked on their right edge with a *particle*[5] corresponding to their grammatical roles within the sentence; for example, as case markers, parallel markers or binding markers (Martin, 2004). Among these is the topic-marking particle *wa*.

In many sentences, *wa* can be used interchangeably with the nominative/subject marker *ga*, as the *topic* of a sentence is often the same as its *subject*.

(10) *Hanako {ga/wa} konakatta*
Hanako {NOM/TOP} come-NEG-PST

'Hanako didn't come.'

A *wa*-marked consituent my also appear as a topic in double-subject constructions (Kumashiro and Langacker, 2003), where a main constituent is the subject but an outer subject is also required, as in (11) where the subject of the adjective *suki* is *Takashi*, but the thematic *wa* indicates the semantic meaning that Hanako is the agent liked *by Hanako*.

(11) *Hanako wa Takashi ga suki*
Hanako TOP Takashi NOM liked(ADJ)

'Hanako likes Takashi.' (lit: 'As for Hanako, Takashi is liked [by her]'

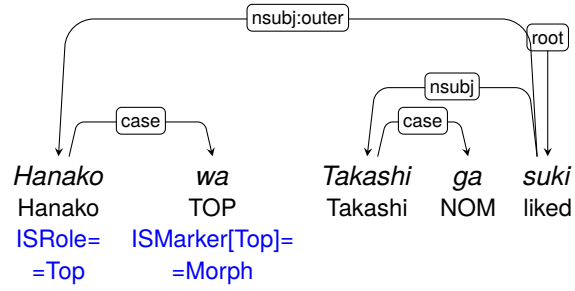
(10) and (11) are unmarked; there is nothing unusual in Japanese about a subject being topicalised, nor a topic being used as an outer subject. In a more marked use, *wa* may also be used to topicalise other arguments, such as direct and indirect objects. (12a) shows an example of a semantic object being marked as a topic, while (12b) shows the more typical accusative-marked type. Obliques may also be made topics through the use of *wa*, as in (13).[6]

(13) *Ôsaka (ni) wa ippai aru*
Osaka (LOC) TOP full exist

'There's loads of it in Osaka.' (lit: 'As for (in) Osaka, loads [of it] exist.')

The exact function of this particle in terms of Information Structure is debated. Oshima (2009) posits that it is an amalgam of a topic-marker and a ground marker, with different roles depending on what it marks. In the case of a *wa*-marked subject, this merely indicates that the subject is part of the ground, but not necessarily a topic; in the case of a *wa*-marked direct object, it indicates that the object is both part of the background and a topic. We adopt the treatment of it as a topic marker, mindful that this may overgenerate topic information.

Japanese as annotated in Universal Dependencies tokenises at the morpheme level and does not attach any morphological features to particles (Tanaka et al., 2016), and thus particles are simply separate tokens tagged as ADP, with a *case* relation to their head, and no additional morphology. For our purposes, this means that we can add the `ISMarker[Top]` attribute to the particle token itself, with the value of `Morph`, while the `ISRole` is added to the head of the phrase.



There are other Information Structure-marking strategies in Japanese that are worth consideration, such as scrambling (Maki et al., 1999) and right-dislocation (Takita, 2014). Maki et al. posit that scrambling in particular may have effects on the role played by *wa*, changing the phrasal unit that it attaches to from a simple topic to a contrastive focus, in which case we may use the `CFoc` value for `ISRole`. We do not commit to this yet, and leave it for further deliberation.

[5] Particle is the term typically used in Japanese linguistics, but in Universal Dependencies they are generally tagged as ADP.

[6] Note that in the oblique (13), a locative marker *ni* is also used before *wa*, while the object in (12a) the object marker *o* is not used before *wa*, and would be ungrammatical if it were.

6.4. Hungarian (Uralic)

In Hungarian, focus expressing exhaustive identification occupies a dedicated position within the predicate phrase: it must appear immediately before the finite verb (É Kiss, 2002; É Kiss et al., 2003).

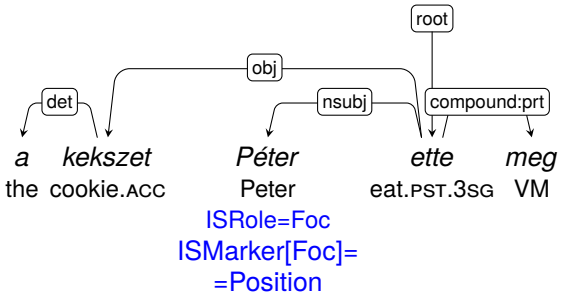
A clear syntactic diagnostic for exhaustive identification comes from the behaviour of so-called verb modifiers (VM), which reveal that focus occupies a dedicated preverbal position. In the Hungarian grammatical tradition, verb modifiers are treated as adverbial elements. They modify the lexical or aspectual meaning of the verb, often contributing telicity, directionality, or other Aktionsart-related distinctions (É Kiss et al., 2003). In neutral clauses, the predicate phrase typically begins with a verb modifier followed by the verb (VM–V order), see (14).

(14) *A kekszet Péter meg-ette.*
the cookie.ACC Peter VM=eat.PST.3SG
'Peter has eaten the biscuit.'

Although the verb and its modifier may form a lexical and phonological unit when adjacent, the modifier is syntactically independent and can be separated from the verb. We therefore analyse it as a clitic (rather than an affix, as in traditional analyses). This independence is clearly visible in focus constructions, where a constituent of the clause bearing exhaustive focus must occupy the immediate preverbal position. Since this position can host only one phrasal unit, the verb modifier cannot remain there and is displaced to a postverbal position. The alternation between neutral VM–V order and focus-induced Foc–V–VM order thus reflects the structural requirement that exhaustive focus be realised in the immediate preverbal slot, see (15) and (É Kiss et al., 2003; É Kiss, 2002).

(15) *A kekszet Péter ette meg.*
the cookie.ACC Peter eat.PST.3sg VM
'PETER has eaten the biscuit.'

In our annotation scheme, such focused units are marked as `ISRole=Foc` and `ISMarker[Foc]=Position`, indicating that focus is encoded configurationally. An example of this analysis is provided in the following UD tree.[7]



It should also be noted that in Hungarian, topics occupy a dedicated syntactic position at the left edge of the clause (É Kiss et al., 2003; É Kiss, 2002). For reasons of space, we do not discuss this construction in detail here. In such cases, however, the topicalised constituent is annotated with `ISRole=Top`, and `ISMarker[Top]` is likewise set to `Position`.

6.5. Italian (IE – Romance)

In Italian, topic and focus are not morphologically encoded. Rather, these information-structural functions are realised through syntactic role, constituent order, and prosody (Benincà et al., 1988; Palermo, 2013; Lombardi Vallauri, 2009, 2014).[8]

More specifically, under canonical word order, corresponding to the (S)VO pattern in Italian, the topic is typically associated with subject function, whereas the focus is commonly realised in object position. Constituent order also plays a central role, as the topical element generally occupies a left-peripheral position (Palermo, 2013). In example (16), the topic corresponds to the subject *Carlo*, which occupies the left-peripheral position, while the focus coincides with the direct object *la lettera* 'the letter', which appears in rightmost position.

(16) *Carlo ha spedito la lettera*
Carlo have.PRS.3SG mail.PTCP the letter
'Carlo has mailed the letter.'

To signal a misalignment with respect to canonical constituent order, Italian employs marked constructions. Given the wide range of such constructions, only two types are presented here, one per category, for illustrative purposes.

Among topic-marking constructions is left dislocation (De Santis and Prandi, 2019), so termed because a constituent is displaced to the left periphery and placed outside the clause core; in written

[7] Since verb modifiers in Hungarian behave similarly to compounds in several respects, they have been analysed as `compound:preverb` in the Hungarian treebank. We prefer to use the `compound:prt` relation for this particle, analogously to how such particles are treated in German and English treebanks.

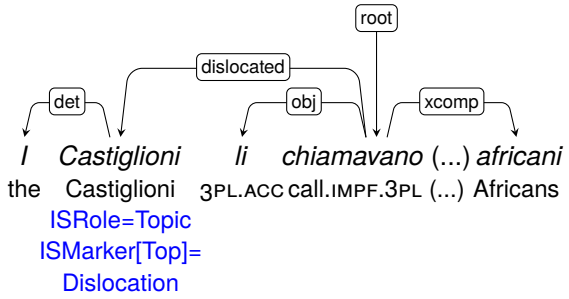
[8] An annotation work on the Information Structure of spoken Italian based on Informational Patterning Theory (Cresti, 2000) is represented by the C-ORAL-ROM corpus (Cresti and Moneglia, 2005); see also (Cresti and Moneglia, 2010).

texts, it is typically set off by a comma. The dislocated constituent is resumed within the clause by a coreferential resumptive pronoun[9].

(17) *I Castiglioni li chiamavano per*
the Castiglioni ACC.3PL call.IMPF.3PL for
scherzo "gli africani"
joke "the africans"

'They used to call the Castiglioni "the Africans" as a joke.' (isst_tanl-404)

In Example (17), *I Castiglioni* 'the Castiglioni' is left-dislocated and functions as the topic of the sentence. Along with the dependency relation `dislocated`, it is marked as `ISRole=Topic` and labeled with `ISMarker[Top]=Dislocation`, as shown in the UD tree.

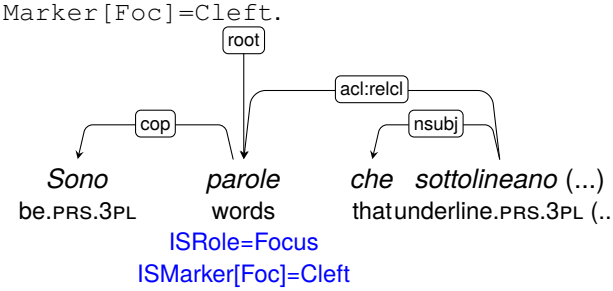


Among the strategies used in Italian to encode focus is the cleft sentence construction (Panunzi, 2009), whereby the unmarked sentence is divided into two clauses: the first is characterised by the verb *essere* 'to be' and the focused element, while the second, introduced by *che* 'that', contains the remainder of the sentence. Cleft sentences employ both intonational and syntactic marking.

(18) *Sono parole che sottolineano (...)*
be.PRS.3PL words that underline.PRS.3PL (...)

'They are words that underline (...)' (markit-945)

In Example (18), the clefted element *parole* 'words' is marked as `ISRole=Focus` and specifies the construction, namely via the label `ISMarker[Foc]=Cleft`.



[9]However, the presence of the resumptive clitic is not always obligatory, especially when the dislocated constituent does not correspond to a direct object (Palermo, 2013).

7. Future work

The immediate next step in this project is to finalise semi-definitive annotation guidelines, providing a clear framework for annotators while allowing for future refinement. Our current focus has been on elements that are overtly marked for Information Structure. Moving forward, we will also annotate inherently marked elements that can be reliably extracted from the lexicon, such as operators and other IS-triggering items.

These guidelines will then be applied to selected UD treebanks. We will prioritise languages where the strategies for marking Information Structure roles are reasonably well understood, while striving to ensure that the sample represents a broad spectrum of morphosyntactic marking strategies.

Once annotation is underway, we will compare overtly marked versus inherently marked elements to identify additional morphosyntactic patterns and correlations. These findings will allow us to formulate generalisations about the interplay between marking strategies and Information Structure roles. Finally, we will situate our results within the broader context of existing approaches, comparing patterns and outcomes to evaluate consistency and typological relevance.

8. Conclusion

We have presented a proposal for annotating Information Structure within the Universal Dependencies framework that preserves the integrity of the existing UD format while providing explicit annotation of constructions and elements that clearly mark Information Structure roles. Our scheme is deliberately minimal and modular, focusing on overtly marked elements as well as inherently marked lexical triggers where they can be reliably identified. This approach allows for consistent annotation across languages without introducing speculative or discourse-specific interpretations.

We illustrated the scheme with case studies from five typologically diverse languages – Wolof, Japanese, Tundra Nenets, Hungarian, and Italian – demonstrating how the framework can capture cross-linguistic variation in Topic, Focus, and Contrast marking strategies. While the scheme is currently tentative, it provides a practical foundation for systematic annotation and comparison, and can be refined or expanded as further data and insights become available.

By anchoring Information Structure annotation to overtly marked elements and clear morphosyntactic triggers, this framework provides a resource for both theoretical linguists and UD users interested in the interaction of discourse and syntax. Beyond immediate annotation, it lays the ground-

work for empirical investigations into the correlation between syntactic structure and discourse roles, the typology of marking strategies, and the development of automated tools for Information Structure analysis.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

Ethical considerations

The authors are unaware of any ethical issues arising from this project.

References

Bibliographical References

Paola Benincà, Giampaolo Salvi, and Lorenza Frison. 1988. L'ordine degli elementi della frase e le costruzioni marcate. In Lorenzo Renzi, editor, *Grande grammatica italiana di consultazione*, volume 1, pages 115–225. Il Mulino, Bologna.

Bernd Bohnet, Alicia Burga, and Leo Wanner. 2013. Towards the annotation of Penn TreeBank with information structure. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1250–1256, Nagoya, Japan. Asian Federation of Natural Language Processing.

Olivier Bondéelle and Sylvain Kahane. 2020. Subjecthood and annotation: The cases of french and wolof. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW)*, pages 34–45.

Olivier Bondéelle and Sylvain Kahane. 2021. Les particules verbales du wolof et leur combinatoire syntaxique et topologique. *Bulletin de la Société de Linguistique de Paris*, 115(1):391–465.

Hannah Booth. 2022. Desiderata for the annotation of information structure in complex sentences. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 31–43, Marseille, France. European Language Resources Association.

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and topic*. Academic Press.

Emanuela Cresti. 2000. *Corpus di Italiano Parlato*. Accademia della Crusca, Firenze.

Emanuela Cresti and Massimo Moneglia, editors. 2005. *C-ORAL-ROM. Integrated reference corpora for spoken Romance languages*. John Benjamins, Amsterdam.

Emanuela Cresti and Massimo Moneglia. 2010. Informational patterning theory and the corpus-based description of spoken language. the compositionality issue in the topic-comment pattern. In Massimo Moneglia and Alessandro Panunzi, editors, *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze University Press, Firenze.

Mary Dalrymple and Irina Nikolaeva. 2011. *Objects and information structure*. Cambridge University Press.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Cristiana De Santis and Michele Prandi. 2019. *Manuale di linguistica e di grammatica italiana*. UTET.

Cheikh M. Bamba Dione. 2019. Developing universal dependencies for wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW), SyntaxFest*, pages 12–23.

Stefanie Dipper, Michael Götze, and Stavros Skopeteas. 2007. Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax. *Interdisciplinary Studies on Information Structure: Working Papers of the SFB632*.

Katalin É Kiss. 2002. *The syntax of Hungarian*. Cambridge University Press.

Katalin É Kiss, Péter Siptár, and Ferenc Kiefer. 2003. Új magyar nyelvtan.

Caroline Fery and Manfred Krifka. 2008. Information structure: Notional distinctions, ways of expression. *Unity and Diversity of Languages*, pages 123–136.

Jan Firbas. 1964. On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague*, 1:267–280.

Jan Firbas. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press, Cambridge.

Jeanette K Gundel. 1988. Universals of topic-comment structure. *Studies in syntactic typology*, 17(1):209–239.

Eva Hajičová, Barbara H. Partee, and Petr Sgall. 1998. *Topic–Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht.

Manfred Krifka. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3-4):243–276.

Toshiyuki Kumashiro and Ronald Langacker. 2003. Double-subject and complex-predicate constructions. *Cognitive Linguistics*, 14(1):1–45.

Knud Lambrecht and Maria Polinsky. 1997. Typological variation in sentence-focus constructions. *Chicago Linguistic Society*, 33(2):189–206.

Edoardo Lombardi Vallauri. 2009. *La struttura informativa. Forma e funzione negli enunciati linguistici*. Carocci, Roma.

Edoardo Lombardi Vallauri. 2014. The topologic hypothesis of prominence as a cue to information structure in italian. In Salvador Pons Bordería, editor, *Discourse Segmentation in Romance Languages*, pages 219–241. John Benjamins, Amsterdam and Philadelphia.

Anke Lüdeling, Julia Ritz, Manfred Stede, and Amir Zeldes. 2016. Corpus linguistics and information structure research. In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*, pages 599–617. Oxford University Press, Oxford, UK.

Hideki Maki, Lizanne Kaiser, and Masao Ochi. 1999. Embedded topicalization in english and japanese. *Lingua*, 109:1–14.

Samuel Elmo Martin. 2004. *A Reference Grammar of Japanese*, chapter 2. University of Hawai'i Press.

Igor Mel'čuk. 2001. *Communicative organization in natural language*. John Benjamins Publishing Company.

Jiří Mírovský, Kateřina Rysová, Magdaléna Rysová, and Eva Hajičová. 2013. (pre-)annotation of topic-focus articulation in Prague Czech-English Dependency Treebank. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 55–63, Nagoya, Japan. Asian Federation of Natural Language Processing.

Ad Neeleman, Elena Titov, Hans Van de Koot, Reiko Vermeulen, et al. 2009. A syntactic typology of topic, focus and contrast. *Alternatives to cartography*, 1551(10.1515):9783110217124–15.

Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*. Walter de Gruyter GmbH & Co KG.

David Y. Oshima. 2009. On the So-Called Thematic Use of Wa: Reconsideration and Reconciliation. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 405–414, Hong Kong. City University of Hong Kong.

Patrizia Paggio. 2006. Annotating information structure in a corpus of spoken danish.

Massimo Palermo. 2013. *Linguistica testuale dell'italiano*. Il mulino.

Alessandro Panunzi. 2009. *Strutture scisse e pseudoscisse : valori d'uso del verbo essere e articolazione dell'informazione nell'italiano parlato*. Franco Cesati Editore, Firenze :.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: an evaluation across different types of texts. In *LREC*.

Stéphane Robert. 1991. *Approche énonciative du système verbal: le cas du wolof*. CNRS Editions.

Stéphane Robert. 2000. Le verbe wolof ou la grammaticalisation du focus. *Topicalisation et focalisation dans les langues africaines*, pages 229–267.

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Serge Sauvageot. 1965. *Description synchronique d'un dialecte wolof : le parler du Dyolof*. Ph.D. thesis, Paris.

Sanghoun Song. 2017. *Modeling information structure in a cross-linguistic perspective*. Language Science Press, Berlin.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Manfred Stede and Sara Mamprin. 2016. Information structure in the Potsdam commentary corpus: Topics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1718–1723, Portorož, Slovenia. European Language Resources Association (ELRA).

Kensuke Takita. 2014. Pseudo-right dislocation, the bare-topic construction, and hanging topic constructions. *Lingua*, 140:137–157.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association (ELRA).

William H. Torrence. 2005. *On the distribution of complementizers in Wolof*. Ph.D. thesis, University of California, Los Angeles.

Henri Weil. 1879. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes: question de grammaire général*, volume 18. F. Vieweg, Paris.

Malte Zimmermann and Caroline Féry. 2010. *Information structure: Theoretical, typological, and experimental perspectives*. Oxford University Press, USA.

Katalin É. Kiss. 1998. Identificational focus versus information focus. *Language*, 74(2):245–273.

Language Resource References

Bejček, Eduard and Hajičová, Eva and Hajič, Jan and Jínová, Pavlína and Kettnerová, Václava and Kolářová, Veronika and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Panevová, Jarmila and Poláková, Lucie and Ševčíková, Magda and Štěpánek, Jan and Zikánová, Šárka. 2013. *Prague Dependency Treebank 3.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antònia Martí, Marie Mikulová, Kirill Milintsevich, Vandan Mujadia, Judith Muzerelle, Sangha Nam, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Ian Porada, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Daniel Swanson, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).