Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 31–43
Marseille, 24 June 2022

Desiderata for the Annotation of Information Structure in Complex Sentences

Hannah Booth
Ghent University
Department of Linguistics, Blandijnberg 2, 9000 Gent, Belgium
hannah.booth@ugent.be

Abstract

Many annotation schemes for information structure have been developed in recent years (Calhoun et al., 2005; Paggio, 2006; Götze et al., 2007; Bohnet et al., 2013; Riester et al., 2018), in line with increased attention on the interaction between discourse and other linguistic dimensions (e.g. syntax, semantics, prosody). However, a crucial issue which existing schemes either gloss over, or propose only crude guidelines for, is how to annotate information structure in complex sentences. This unsatisfactory treatment is unsurprising given that theoretical work on information structure has traditionally neglected its status in dependent clauses. In this paper, I evaluate the status of pre-existing annotation schemes in relation to this vexed issue, and outline certain desiderata as a foundation for novel, more nuanced approaches, informed by state-of-the art theoretical insights (Erteschik-Shir, 2007; Bianchi and Frascarelli, 2010; Lahousse, 2010; Ebert et al., 2014; Matić et al., 2014; Lahousse, 2022). These desiderata relate both to annotation formats and the annotation process. The practical implications of these desiderata are illustrated via a test case using the Corpus of Historical Low German (Booth et al., 2020). The paper overall showcases the benefits which result from a free exchange between linguistic annotation models and theoretical research.

Keywords: annotation, information structure, complex sentences, subordination, historical data, Middle Low German

1. Introduction

Recent years have seen a boom in language resources which contain some form of information-structural (IS) annotation, for which various schemes and guidelines have been developed (Calhoun et al., 2005; Paggio, 2006; Götze et al., 2007; Bohnet et al., 2013; Riester et al., 2018). However, the issue of dependent clauses for IS annotation has been largely neglected; many have acknowledged complex sentences as an annotation challenge for IS (Bohnet et al., 2013; Cook and Bildhauer, 2013; Stede and Mamprin, 2016), but few efforts have been made to get to grips with the issue in a concrete and nuanced way. Moreover, theoretical work has highlighted the special status of dependent clauses with respect to IS and related interface phenomena, and thus suggests that we disregard this aspect of IS annotation at our peril (Hooper and Thompson, 1973; Haiman, 1978; Bybee, 2002; Bianchi and Frascarelli, 2010; Lahousse, 2010; Ebert et al., 2014; Matić et al., 2014; Lahousse, 2022).

Neglect of this issue can result in inaccurate and/or conflicting annotations, or even unannotated data. Such outcomes are unsatisfactory and hold back research progress, both theoretical and computational. Without a proper treatment of IS in dependent clauses, theoretical research into the discourse properties of complex sentences and how this interacts with e.g. morphosyntactic and prosodic phenomena cannot rely on the types of corpus-based, quantitative and reproducible investigations which have proven so fruitful in other domains of linguistics. Computational research is also disadvantaged in this context, as inaccurate, conflicting or absent IS annotations, even if confined to a subset of contexts, will inevitably impact NLP downstream tasks.

In this paper, I respond to this challenge by outlining desiderata for the annotation of IS in complex sentences, which can serve as a foundation for novel and nuanced approaches in future. These proposals are underpinned by theoretical insights and are also informed by previous IS annotation schemes which have highlighted specific problems concerning complex sentences. The desiderata relate to aspects of both the annotation format and the annotation process, and are tested in relation to the IS annotation of Middle Low German texts (*c.* 1200–1650) in the Corpus of Historical Low German (Booth et al., 2020), which are known to exhibit highly complex sentence structures (Tophinke, 2012).

2. Theoretical Insights

The IS properties of complex sentences constitute a highly relevant though understudied domain (Matić et al., 2014). Moreover, even from the existing literature on the matter, it is hard to establish a general consensus on even essential questions. This lack of consensus is particularly problematic in the context of linguistic annotation, where schemes which are as theoretically neutral as possible and compatible with different approaches are seen as the gold standard (Bird and Liberman, 2001; Ide and Romary, 2004). In this section, I discuss to what extent some common ground can be established from previous discussions of IS in complex sentences, highlighting crosslinguistic generalisations as well as matters which require nuanced treatment.

2.1. Information-Structural Primitives

A range of theoretical approaches to IS have emerged over recent decades and views differ as to the precise primitives involved and their diagnostic criteria; for useful overviews see e.g. Vallduví (1992); von Heusinger (1999); Büring (2007); de Swart and de Hoop (2014). This paper mainly discusses topic and focus. I follow approaches where topichood is understood as comprising (i) A(BOUTNESS)-TOPIC, (cf. "sentence topic", Reinhart (1981; Krifka (2007)) and (ii) F(RAME)-TOPIC (Krifka, 2007), as defined in (1). Focus is understood as covering (i) I(NFORMATIONAL)-FOCUS (Reinhart, 1981; Vallduví, 1992) and (ii) C(ONTRASTIVE)-FOCUS (Neeleman et al., 2009), cf. (2).[1]

(1) **Topic**
- A(BOUTNESS)-TOPIC: entity/proposition about which a main clause predicates
- F(RAME)-TOPIC: frame within which the main clause predication is interpreted

(2) **Focus**
- I(NFORMATIONAL)-FOCUS: new info which is most relevant to current discourse
- C(ONTRASTIVE)-FOCUS: element/proposition which evokes alternatives

Additionally, I discuss COMMENT, i.e. what is said about the topic, and BACKGROUND, which is material which is neither topic nor focus.

2.2. The Domain(s) of Information Structure

A central issue on which views differ concerns what the precise domain(s) of IS is/are, or more specifically, to what extent dependent clauses can be considered to have IS articulation(s) in their own right. The traditional view is that the domain of IS is the overall utterance, i.e. that even a complex sentence has IS articulation(s) only at the matrix level (Mathesius, 1975; Vallduví, 1992; Vallduví and Zacharski, 1994; Steedman, 2000; Komagata, 2003). However, more recent work assumes that IS can operate within a single utterance at different levels, allowing for dependent clauses to be considered as a potential IS domain. In particular, the notion of recursive IS has been adopted by many (Koktová, 1996; Partee, 1996; Hajicová et al., 1998; Erteschik-Shir, 2007; Matić et al., 2014), with a distinction between (i) "external IS", i.e. the IS status of a dependent clause in the overall matrix clause and (ii) "internal IS", i.e. the IS status of individual constituents within a dependent clause (Erteschik-Shir, 2007; Matić et al., 2014). These two perspectives are illustrated in (3) and (4) respectively (Matić et al., 2014, 9-10). In (3) (external IS), the whole matrix sentence is considered as the relevant IS domain, in which the clefted adverbial clause *after I arrived home* is assigned focus. In (4) (internal IS), the complement clause is viewed as an IS domain its own right, within which *this book* receives a topical interpretation.

(3) [It was only $\underbrace{\textbf{after I arrived home}}_{\text{FOCUS}}$ that I saw them].

(4) I believe [that $\underbrace{\textbf{this book}}_{\text{TOPIC}}$ Mary gave to Paul].

Combining these two perspectives yields recursion, whereby a dependent clause can be a topic/focus with respect to external IS, but can also contain an internal topic/focus, e.g. (5) and (6) (Partee, 1996, 79, 82).

(5) [**What convinced Susan that [our arrest]$_{\text{TOPIC}}$ was caused by Harry**]$_{\text{TOPIC}}$ was a rumour that someone had witnessed Harry's confession.

(6) What convinced Susan that our arrest was caused by Harry was [**a rumour that someone had [witnessed Harry's confession**.]$_{\text{FOCUS}}$]$_{\text{FOCUS}}$

In line with the majority of recent work, I assume that dependent clauses can in principle have internal IS articulation(s) under certain conditions, as I discuss next.

2.3. Assertion and Clause Class

It is widely recognised that the possibility of a clause having internal IS is connected with assertion; clauses which are asserted are more likely to have internal IS than clauses which are presupposed (Bybee, 2002; Lahousse and Borremans, 2014; Matić et al., 2014). Dependent clauses are traditionally understood as being presupposed rather than asserted (Quirk et al., 1985; Hooper and Thompson, 1973; Matsuda, 1998), and thus less susceptible to internal IS permutations (Lehmann, 1988; Bybee, 2002). However, general distinctions can be drawn between different classes of dependent clause, and indeed even within some classes. Complement clauses, for instance, are more likely to have internal IS than adverbial and relative clauses, since the former are often asserted and the latter typically presupposed (Matić et al., 2014).

At the same time, a long-standing body of research has shown that the internal IS of complement clauses is conditioned by the type of embedding predicate in the matrix clause. Only complement clauses which represent the main assertive point, i.e. are embedded under nonfactive predicates, can have an articulated internal IS (Matić et al., 2014), in line with observations that phenomena connected with topicality are restricted to such contexts (Hooper and Thompson, 1973; Boye and Harder, 2007; Dehé and Wichmann, 2010; Matić et al., 2014). For instance, English topic marking via fronting is permitted in the complement of the nonfactive predicate *explain* in (7) (Hooper and Thompson, 1973, 474) but ruled out under a factive predicate like *regret*, e.g. (8) (Maki et al., 1999, 3).

(7) The inspector explained [that **each part** he had examined very carefully].

[1] In principle also topics can be contrastive, but I do not discuss contrastive topics here.

(8) *John regrets [that **this book** Mary read].

The type of embedding predicate also interacts with the external IS of complement clauses; complements of factive verbs are usually discourse-given and generally unfocable, unless they are contrasted with a competing presupposition (Matić et al., 2014). Complements of nonfactive verbs can however carry the main assertion, and in such cases it has been claimed that the matrix clause is informationally demoted to a parenthetical clause (Dehé and Wichmann, 2010).

Likewise, adverbial clauses do not exhibit consistent IS properties. An important distinction here is between "central" (i.e. event-structuring) and "peripheral" (i.e. discourse-structuring) adverbial clauses (Haegeman, 2007). Central adverbial clauses are more syntactically and prosodically integrated into their host clause than their peripheral counterparts, but they also differ in terms of assertion; the central class is generally assumed to be presupposed, and the peripheral class asserted (Lahousse and Borremans, 2014), which has been used to argue for the peripheral type having internal IS and to explain the occurrence of root-like phenomena in such environments (De Cat, 2012).

Relative clauses also exhibit diverse IS properties, in particular, between nonrestrictive, e.g. (9) and restrictive relative clauses, e.g. (10) (Fabb, 1990, 57).[2]

(9) The swans, **which are white**, are in that part of the lake

(10) The swans **which are white** are in that part of the lake.

With respect to external IS, nonrestrictive relative clauses have been argued to be neither focus nor topic but rather backgrounded (Umbach, 2006; Song, 2014), since they provide extra information about a referent already determined on independent grounds (Riester, 2009). Restrictive relatives provide a description which uniquely identifies a referent, and show many similarities with classic focus constructions such as clefts (Schachter, 1973). With respect to internal IS, restrictive relatives are assumed to lack internal IS (Depraetere, 1996; Matić et al., 2014), since they provide a description which uniquely identifies a referent and must thus contain material which is already part of the "common ground" (Stalnaker, 2002). Nonrestrictive relatives contain new, asserted information and are thus more likely constitute an independent IS domain in their own right (Depraetere, 1996; Bybee, 2002).

2.4. Clause Ordering

The relative ordering of a main clause and its dependent clause(s) often affects their IS relations with each other and the wider discourse (Lehmann, 1988; Diessel, 2001; Schilder and Tenbrink, 2002; Komagata, 2003). In terms of external IS, it has been observed for many languages that dependent clauses which occur before their host clause are often topical (Marchese, 1977; Lehmann, 1984; Thompson, 1985; Chafe, 1984; Lehmann, 1988; Diessel, 2001). Conditional clauses, for instance, which typically occur before the host clause, have been observed to be often topics (Schiffrin, 1992; Ebert et al., 2014), to the extent that this has been claimed to be a universal (Haiman, 1978). Further evidence for the correlation between initial dependent clauses and topicality comes from various languages where initial adverbial clauses are marked by the same morpheme as clause-internal topics (Thompson and Longacre, 1985). An example is Lisu (Tibeto-Burman), where initial adverbial clauses are marked by *nya*, which can also mark a topic in the following main clause, e.g. (11) (Thompson and Longacre, 1985, 232).

(11)

[ame	thæ	nwu	patsɨ-a	dye-a̺	ŋu	
yesterday	TIME	you	plain-to	go-DECL	FACT	
bæ̺-a̺	**nya**]	nwu	**nya**	asa	ma	mu-a.
say-DECL	TOPIC	you	TOPIC	Asa	not	see-Q

'When you went to the plain yesterday, didn't you see Asa?'

Clause ordering has also been shown to be relevant for the internal IS of dependent clauses. Komagata (2003), for instance, claims for English that dependent clauses with their own internal IS only appear after the main clause; dependent clauses which precede a main clause are expected to lack internal IS, in line with the fact that they do not involve assertion but instead relay information already part of the common ground (Lelandais and Ferré, 2017).

3. Previous IS Annotation Schemes

With respect to the treatment of complex sentences, reports on previous IS annotation schemes typically sidestep the issue or propose only a few crude guidelines. For instance, in Buráňová et al. (2000), Baumann et al. (2004) and Calhoun et al. (2005) there are no specific comments regarding the annotation of complex sentences. Elsewhere, a certain amount of attention is given to whether dependent clauses should be treated as having their own internal IS. The guidelines by Paggio (2006), for example, allow dependent clauses to be treated either as an independent IS domain with its own focus and potentially topic, or as simply serving an IS role in the matrix sentence, either as background or part of the focus domain. This is a heuristic used to guide annotation which largely "relies on the coder's intuition" (Paggio, 2006, 1606).

Likewise, in the (otherwise detailed) scheme outlined by Götze et al. (2007), relatively scant detail is provided regarding complex sentences. In terms of topic annotation, they suggest a strategy whereby one first checks whether the whole matrix sentence has an aboutness and/or frame topic. One then examines each finite clause within the complex sentence – with the

[2] In (9), the implication is that all swans under discussion are white; (10) instead implies that the white swans are distinguished from some other swans under discussion.

exception of restrictive relative clauses – to check for whether it has its own aboutness/frame topic. Thus, apart from sidelining restrictive relative clauses, which can be assumed to lack internal IS (see Section 2), no further distinction is made between different classes of dependent clause.

In subsequent tests of Götze et al.'s guidelines for topic annotation (Cook and Bildhauer, 2011; Cook and Bildhauer, 2013), complex sentences were found to be a problematic area for annotation consistency. A particular challenge was whether to annotate dependent clauses for internal IS, and whether different embedding predicates/clause classes merit different approaches. On this point, Stede and Mamprin (2016) include some revisions to Götze et al.'s guidelines, limiting topic annotation to adverbial clauses and excluding complement clauses. This though is an oversimplistic generalisation, which does not acknowledge that internal topics are possible in complement clauses embedded under certain predicates, cf. (7) above.

Bohnet et al. (2013), who assume a tripartite IS articulation ("Theme-Rheme-Specifier"), allow for recursive IS; if a dependent clause constitutes its own proposition, it can be annotated in terms of both external and internal IS.[3] An example is shown in (12) (Bohnet et al., 2013, 1251), where the relative clause belongs both to the R(heme) of the matrix sentence but is itself segmented into T(heme) and (R)heme.

(12) [Years ago]$_{\mathrm{SP}}$, [he]$_{\mathrm{T}}$ [collaborated with the new music gurus Peter Serkin and Fred Sherry in the very countercultural chamber group Tashi, [**which**]$_{\mathrm{T}}$ [**won audiences over to dreaded contemporary scores like Messiaen's Quartet for the End of Time**]$_{\mathrm{R}}$]$_{\mathrm{R}}$.

Nonetheless, Bohnet et al. (2013) acknowledge that in highly complex sentences, their parser for automatic thematicity annotation suffers errors arising from the incorrect detection of the propositions involved.

Riester et al. (2018) also address the question of what constitutes an IS domain in their Question-Under-Discussion (QUD) approach to IS annotation (von Stutterheim and Klein, 1989; van Kuppevelt, 1995) . With respect to dependent clauses, they rely on *at-issueness* as a diagnostic. Non-at-issue content, i.e. content which does not answer the current QUD, expressed by adverbial and nonrestrictive relative clauses, is treated as lacking internal IS.

In sum, the main challenges highlighted within preexisting IS annotation schemes include (i) to what extent dependent clauses should be annotated for internal IS, and (ii) whether generalisations can be assumed and employed for the IS properties of different classes of dependent clause.

4. Desiderata for IS Annotation in Complex Sentences

In this section, I outline certain desiderata which can inform future, more nuanced schemes for the annotation of IS in complex sentences, in line with the theoretical insights discussed in Section 2 and the practical issues identified for previous schemes in Section 3. Some of these desiderata derive from the general nature of IS itself, but many are motivated by the specific issues which complex sentences raise. Language-specific concerns are expected, but here I concentrate on the crosslinguistic generalisations which can be drawn. I distinguish between desiderata which relate to (i) annotation format and (ii) the annotation process.

4.1. Annotation Format

While IS annotation can in principle span a range of different formats, one can nevertheless identify certain key features which any chosen format should be able to handle, in order to achieve a theoretically sound and practically sensible IS annotation: (i) multiplicity, (ii) recursion, (iii) discontinuity, (iv) supra-clausality, (v) uncertainty and (vi) meta-annotation.

4.1.1. Multiplicity

Even at the matrix level alone, any IS annotation scheme needs to be able to handle multiplicity, i.e. multiple, potentially cross-cutting IS articulations within a single clause/sentence. Firstly, it is generally acknowledged that topic and focus are not evaluated on the same basis, and as such cannot be considered complements of one another (Vallduví, 1992; von Heusinger, 1999; de Swart and de Hoop, 2014). As such, topic-comment and focus-background articulations cross-cut each other in various ways. A classic example is provided by Dahl (1974), repeated here in (13) (as discussed by Vallduví (1992, 55)).

(13) Q: What does John drink?

A$_1$: $\underbrace{\text{John}}_{\text{TOPIC}}\ \underbrace{\text{drinks beer}}_{\text{COMMENT}}$

A$_2$: $\underbrace{\text{John drinks}}_{\text{BACKGROUND}}\ \underbrace{\text{beer}}_{\text{FOCUS}}$

Multiplicity can also surface in clauses which contain multiple topics/foci, although this is a controversial area (Erteschik-Shir, 2007). Some have argued that a clause can contain more than one aboutness topic (Nikolaeva, 2001; Erteschik-Shir, 2007; Krifka and Musan, 2012; Dalrymple and Nikolaeva, 2011), in particular when a relation between two entities is expressed and commented on, e.g. (14) (Krifka and Musan, 2012, 29). Many languages have also been argued to exhibit multiple foci (Krifka, 2007; Surányi, 2007; Hedberg, 2013), e.g. (15) (Krifka, 2007, 258).

(14) As for **Jack**$_{\mathrm{TOPIC}}$ and **Jill**$_{\mathrm{TOPIC}}$, they married last year.

(15) John only introduced **Bill**$_{\mathrm{FOCUS}}$ only to **Sue**$_{\mathrm{FOCUS}}$.

[3] Theme and Rheme are roughly equivalent with (aboutness) topic and comment,; the Specifier sets of the context of the utterance (≈ frame topic).

4.1.2. Recursion

The issue of recursion presented in particular by dependent clauses is a different type of challenge, cf. (12) above. This ultimately requires some level of hierarchisation in a single annotation layer. Hierarchical structure is no stranger to linguistic annotation, being widely employed in e.g. syntactic annotation schemes which encode constituency (Brants et al., 2002; Taylor et al., 2003). However, the majority of the previous IS annotation schemes encode IS via flat spans. Moreover, since many IS annotation contexts involve adding IS annotations to a syntactically annotated resource, further hierarchical IS annotations must be carefully designed so as not to result in conflicting hierarchies.

4.1.3. Discontinuity

Many languages exhibit discontinuous IS fields, i.e. when a single IS status is assigned to multiple non-adjacent segments, e.g. (16) (German), which shows a discontinuous focus (Gussenhoven, 1999, 50), and (17) (Serbian), which shows a discontinuous topic (Milićev and Milićević, 2012, 207).[4]

(16) *What happened to the child?*
Karl hat dem Kind **einen Füller** **geschenkt**
Karl has the child a fountain-pen given
'Karl gave the child a fountain pen'

(17) **Marija** sutra, **profesorica latinskog**, odlazi u penziju.
Mary tomorrow professor of-Latin goes to retirement
'Mary, professor of Latin, retires tomorrow.'

Discontinuous phenomena are of course not limited to IS; at the syntactic level, for instance, much work has focused on the representation of discontinuous constituents in linguistic annotation (Boyd, 2007; Maier and Lichte, 2011), but the issue has generally not been addressed in relation to IS annotation.

4.1.4. Supra-clausality

Another issue which arises in particular in relation to the annotation of complex sentences is the need to encode IS fields which are supra-clausal, i.e. span across clause boundaries. Examples of this were already provided in (5) and (12). This issue is particularly pertinent in contexts where IS annotation is combined with some form of syntactic annotation. The format must allow for IS annotations to cross-cut syntactic clause boundaries. In other words, IS annotation cannot simply be parasitic on syntactic annotation; it must have sufficient autonomy.

4.1.5. Uncertainty

Any IS annotation scheme should also be able to encode some level of uncertainty in contexts where a clear-cut identification of IS domains and/or classification of IS articulations cannot be made. The annotation of uncertainty has attracted attention in recent years (Barteld et al., 2014; Merten and Seemann, 2018; Andresen et al., 2020; Beck et al., 2020), and is particularly critical for IS annotation across complex sentences where our theoretical knowledge is still underdeveloped. In particular, whereas much of the theoretical understanding of IS is formulated on the basis of isolated question-answer pairs, the identification and classification of IS in long stretches of natural linguistic data, where non-directly questionable dependent clauses are commonplace, is less straightforward (Lüdeling et al., 2016).

Uncertainty with respect to IS annotation can arise in relation to two different aspects: (i) whether a particular segment constitutes an independent IS domain with its own internal IS articulation(s) and (ii) how and where the IS articulation(s) in a given IS domain should be drawn. The former is particularly relevant in the context of complex sentences where, as discussed in Section 2, views differ as to whether dependent clauses can be IS domains in their own right. As such, some mechanism for capturing (different types of) uncertainty, ideally based on a relatively sophisticated propagation model like that envisaged by Beck et al. (2020), should be a crucial component of any IS annotation scheme.

4.1.6. Meta-annotation

IS annotation schemes should also have the capability of encoding some form of meta-annotation, i.e. information about a given IS annotation, which explains/justifies the choices made. Meta-annotation is generally recognised as an important enhancement to linguistic annotations (Leech, 2005; Smith et al., 2008) and has been implemented in various resources and schemes (Laprun et al., 2002; Romary et al., 2010). It is particularly relevant in the context of IS, which lacks consensus on key concepts and definitions, in particular in relation to complex sentences. As a result, judgements involved are often less clear-cut and more subjective than at other linguistic levels, even with a carefully operationalised set of diagnostic criteria. The use of meta-annotations here can promote the usability of the resources for theoretical studies, making the decision behind the annotation transparent and allowing the user to reclassify the data if desired. In cases where the annotator is uncertain, as discussed above, meta-annotation can also be an important enhancement, setting out the locus of the uncertainty and allowing it to be potentially resolved at a later date.

4.1.7. Summary

Four of the six requirements discussed here (multiplicity, supra-clausality, uncertainty and meta-annotation) can be easily satisfied by employing a stand-off, multi-dimensional annotation format. Such a format in principle allows for independent, linked annotation lay-

[4] On the distinction between multiple foci and discontinuous focus, see Gussenhoven (1999, 49–50).

ers for modelling (i) multiple cross-cutting IS articulations (ii) IS annotations which are autonomous and not structurally dependent on syntactic annotations, (iii) conflicting annotations for a particular IS articulation across co-existing layers in cases of uncertainty or differing theoretical assumptions, and (iv) meta-annotations to aid transparency and usability. At present, the best possibility is to use some stand-off XML format. This is indeed already recommended by e.g. CLARIN-D,[5] and many others have advocated for this format in recent years (Dipper, 2005; Lüdeling et al., 2016) and employed it specifically for IS annotation (Stede and Mamprin, 2016; Celano, 2019). Moreover, purpose-built infrastructures, such as the interoperable *corpus-tools.org* toolchain (Druskat et al., 2016) which caters for the creation, annotation, query and analysis of multidimensional corpora, mean that such projects are relatively achievable. Yet the full potential on offer for capturing the nuances of IS in complex sentences has yet to be exploited.

At the same time, the issues discussed (in particular multiplicity, discontinuity and recursion) also impose demands on the format of individual annotation layers. For any layer which encodes a certain IS articulation, the structural representation of the annotation needs to go beyond labelled spans over continuous segments of text and must be able to capture the distinction between (i) multiple topics/foci in a single clause and (ii) non-adjacent segments which are assigned a single topic/focus value, potentially via some form of coindexation or linking mechanism. Additionally, in order to allow for recursion in complex sentences, IS annotation layers need to allow for hierarchical relations.

4.2. Annotation Process

Manual IS annotation based on pragmatic context-based judgements alone is a relatively subjective and time-intensive process, especially in relation to complex sentences where, as mentioned, our understanding of IS is generally underdeveloped. Overall, various models for the automatic annotation of IS have been trialed (Hempelmann et al., 2005; Nissim, 2006; Cahill and Riester, 2012; Markert et al., 2012; Rahman and Ng, 2012; Ziai and Meurers, 2018), but automatic annotation for IS is not as reliable as for other tasks (Lüdeling et al., 2016). It generally exploits pre-existing annotations for morphosyntactic and lexical features which approximately correlate with IS properties. Most developments in automatic IS annotation focus on the discourse status of referents (e.g. old/new) (Hempelmann et al., 2005; Nissim, 2006; Cahill and Riester, 2012; Markert et al., 2012; Rahman and Ng, 2012), and these approaches thus exploit nominal features, e.g. weight (pronoun/noun), position (sentence-initial/-final), grammatical function (subject/object) and whether the referent has been previously mentioned or not.

To my knowledge, the possibilities for automatic annotation of IS specifically in relation to complex sentences remain as yet unexplored. Given the fact that certain crosslinguistic syntax-IS correspondences can be identified for dependent clauses (see Section 2), it seems sensible to test to what extent these correspondences can be useful in informing a (potentially automated) rule-based approach to the IS annotation of complex sentences, especially since many contexts for IS annotation involve adding additional annotations on top of pre-existing syntactic annotations. In this section, I outline the basis for such an approach, before testing it in Section 5.

The IS annotation process can be broken down into two key tasks: (i) the identification of IS domains and (ii) the classification of IS articulations within those domains. With respect to complex sentences, I argue that adopting an approach whereby each dependent clause is annotated in two separate stages, with respect to (i) external IS and (ii) internal IS (see Section 2), is most efficient. This is because the classification of a dependent clause in terms of its external IS role, and the decision as to whether it has internal IS, are largely independent of each other and informed by different considerations. In particular, it should be borne in mind that identification of an external IS role for a given dependent clause does not necessarily imply that it has internal IS.

4.2.1. Stage I (External IS)

In terms of the external IS of dependent clauses, the most robust crosslinguistic generalisations which can be identified in the literature are those in (18), where *D* stands for dependent clause, RRC for restrictive relative and NRRC for nonrestrictive relative clause.

(18) **Crosslinguistic syntax-IS correspondences**

- *D* occurs before host clause ≈ TOPIC
- *D* is conditional clause ≈ TOPIC
- *D* is clefted ≈ FOCUS
- *D* is nonfactive complement ≈ FOCUS
- *D* is factive complement ≈ BACKGROUND
- *D* is RRC ≈ FOCUS
- *D* is NRRC ≈ BACKGROUND

The correspondences in (18) are general correlations rather than hard and fast constraints. On the basis of these correspondences, I propose the rule-based algorithm in Figure 1 for the assignment of external IS to dependent clauses (*D*), which exploits syntactic/semantic properties. The top split concerns clause ordering, i.e. whether *D* is before the host clause or in another position. If *D* is before the host clause, it is straightforwardly annotated as topic; if *D* occurs in a different position, a range of annotations are possible, subject to clause class and syntactic/semantic properties (clefting/non-restrictiveness). With respect to the (non)factivity of complement clauses, I refer to the predicate classes in Hooper and Thompson (1973).

[5] https://media.dwds.de/clarin/userguide/text/annotation_aspects.xhtml

```
case  D is before host clause
      external IS := TOPIC
case  D is not before host clause
      if   D is conditional clause then
           external IS := TOPIC
      elif D is clefted then
           external IS := FOCUS
      elif D is complement clause then
           if   D is nonfactive then
                external IS := FOCUS
           else
                external IS := BACKGROUND
      elif D is relative clause then
           if   D is RRC then
                external IS := FOCUS
           else
                external IS := BACKGROUND
      else
           external IS := BACKGROUND
```

Figure 1: Hand-crafted rule-based algorithm for assigning external IS to dependent clauses

4.2.2. Stage II (Internal IS)

Stage II represents a more complex set of tasks, involving the decision as to whether a dependent clause constitutes an IS domain with its own internal IS and, if yes, then classifying any relevant IS articulation(s) within that domain. As discussed in Section 3, the correct identification of IS domains in relation to complex sentences has challenged previous approaches to IS annotation and so I focus on this aspect of the internal IS annotation of dependent clauses.

On the basis of the crosslinguistic tendencies discussed in Section 2, I propose the rule-based algorithm in Figure 2 as a heuristic to aid the decision as to whether a given dependent clause constitutes an IS domain with its own internal IS. Again, this exploits clause ordering as the top split, and then clause classes and subclasses at lower levels. This algorithm can also in principle be combined with information as to whether the dependent clause is asserted or presupposed, as assertive status generally indicates internal IS, and presupposed status lack of internal IS. Here, semantic tests for assertion/presupposition are recommended, of which there are a range in the literature, e.g. the denial and question tests (Hooper and Thompson, 1973; Wiklund et al., 2009) for identifying assertions and the negation test (Kiparsky and Kiparsky, 1970; Hooper, 1975) and the *Hey, wait a minute* test (von Fintel, 2004) for identifying presuppositions. Such tests, however, typically rely on time-intensive judgements and should be considered as a potential supplement to the primarily syntactic-based algorithm in Figure 2, which is designed to exploit pre-annotated morphosyntactic and lexical features as far as possible.

```
case  D is before host clause
      status := no internal IS
case  D is not before host clause
      if   D is adverbial clause then
           if   D is central adverbial clause then
                status := no internal IS
           else
                status := internal IS
      elif D is complement clause then
           if   D is factive then
               status := no internal IS
           else
                status := internal IS
      elif D is relative clause
           if   D is RRC then
                status := no internal IS
           else
                status := internal IS
      else
           status := unknown
```

Figure 2: Hand-crafted rule-based algorithm for deciding whether to assign internal IS to dependent clauses

5. Test Case: Middle Low German

The approaches outlined in Section 4.2 were tested in the IS annotation of dependent clauses in a Middle Low German text from the Corpus of Historical Low German (CHLG) (Booth et al., 2020) specifically the text *Engelhus*, which is a Low German version of Dietrich Engelhus' *Chronica Nova*. The text is an historical chronicle from 1435 CE, and contains 709 clauses annotated as dependent clauses (`IP-SUB`) in the syntactic Penn-style annotation, although some of this number will be embedded conjuncts within a larger coordination structure which can likely be assigned a single external IS tag. Moreover, some of the clauses tagged `IP-SUB` will be dependent clauses which themselves are embedded in dependent clauses, which I do not consider for external or internal IS annotation for the purposes of this paper. Whether such multiply embedded dependent clauses should be annotated for their external IS role in the local dependent clause, or exhibit their own internal IS articulations, I leave open for future consideration.

5.1. Annotation of External IS

All dependent clauses in *Engelhus* were manually annotated for external IS on the basis of contextual pragmatic judgements alone (i.e. irrespective of syntactic and lexical features), using the annotation tool Annotald (Beck et al., 2015). The categories which were annotated were as in (19), largely following the diagnostics provided in Götze et al. (2007) (cf. also (1) and (2) in Section 2.1).

(19) **IS tags**

- TOPIC, which includes:
  - A(BOUTNESS)-TOPIC
  - F(RAME)-TOPIC
- FOCUS, which includes:
  - I(NFORMATIONAL)-FOCUS
  - C(ONTRASTIVE)-FOCUS
- BACKGROUND

A fresh round of (manual) annotation was then performed relying exclusively on the rule-based algorithm in Figure 1 as annotation guidelines, without consideration of the pragmatic context. The result was then compared against the first round of annotations in order to assess the algorithm's accuracy. The overall accuracy of the algorithm, i.e. the number of correctly classified instances of all assignments is 81.6%. The precision and recall for each tag is provided in Table 1.

	P	R	F
TOPIC	.849	.753	.798
FOCUS	.860	.636	.731
BACKGROUND	.704	.884	.783

Table 1: Per tag performance of hand-crafted rule-based algorithm for annotation of external IS

A particularly high number of assignments of the BACKGROUND tag were false positives, the majority of which were in fact foci. The over-assignment of the BACKGROUND tag is not surprising, given that this was used as a catch-all for remaining instances of non-initial dependent clauses, cf. Figure 1. As such, future refinements of the algorithm could include finding extra classes/contexts which are likely to coincide with focus for non-initial dependent clauses.

The algorithm in Figure 1 does not distinguish between different types of topic/focus, cf. (19), as it is designed to be crosslinguistically applicable and was thus informed by only the most robust crosslinguistic generalisations. However, with respect to at least Middle Low German, some further language-specific correlations between syntax and specific types of topic/focus can be observed from the first round of pragmatic, context-based annotations, which may perhaps turn out to be more general correlations. For instance, of the 70 dependent clauses which occur before the host clause in *Engelhus*, 67 of these are topics. However, only two of these qualify as aboutness topics, both free relatives in a left-dislocation/resumption structure, e.g. (20).

(20) [**wor auer Noe henkeyme**]$_i$ **dat**$_i$ vindest u
where however Noah comes-to that find you
hir na ffalech
here after Falech
'Wherever Noah comes to though, that you find hereafter, Falech'

The other sentence-initial clauses which qualify as topics (*n*=65) are frame-topics. These were most commonly adverbial clauses, again in a left-dislocation/resumption structure, e.g. (21), or conditional clauses, e.g. (22).

(21) [**Do lamech was clxxii iar** olt]$_i$ **do**$_i$ ghewan
when Lamech was 172 years old then had
he Noe
he Noah
'When Lamech was 172 years old, then he had Noah'

(22) [**wolde eymant eyn belde nomen myner**] de
wanted someone a picture take my.GEN he
nome ok eyn belde mir pyne
take also a picture my.GEN pain.GEN
'If someone wanted to one of my pictures, they would take also a picture of my pain'

With respect to types of focus (information/contrastive), some additional patterns were observed. The (typically nonfactive) complement clauses annotated as focus were all assigned specifically information focus in terms of their external IS, e.g. (23), whereas restrictive relative clauses were typically annotated as contrastive focus, since their function to uniquely identify a referent implies the presence of alternatives, e.g. (24).

(23) Me schrift von eme [**dat he lachede do**. . .]
one writes of him that he laughed when
'One write of him he laughed when. . . '

(24) it wore de [**de ore gode vorstoren scholde**]
it be.SBJV DEM REL her god destroy should
'unless it were she who was to destroy her god (and not someone else)'

As such, it seems that, for MLG at least, one should acknowledge extra syntax-IS correlations for dependent clauses, which pertain to specific types of topic/focus. Further crosslinguistic research would however need to be conducted before these could be included in the algorithm in Figure 1, which is intended to be crosslinguistically applicable.

5.2. Annotation of Internal IS

In a separate task, each dependent clause in *Engelhus* was manually annotated on the basis of pragmatic judgements alone for the presence/absence of internal IS, on the basis of whether internal IS articulations could be identified given the context, again largely following the guidelines in Götze et al. (2007) for the identification of aboutness/frame topics, information/contrastive foci, cf. (19). Dependent clauses were also explicitly annotated if they lacked internal IS.

A fresh round of (manual) annotation was then performed using the rule-based algorithm in Figure 2 as guidelines to classify each dependent clause as either having or lacking internal IS, without paying attention to the pragmatic context. The results of the algorithm were then compared with the first round of annotations to assess the algorithm's accuracy at identifying internal IS contexts, which is known to be a challenging area in the IS annotation of complex sentences (see Section 3).

Overall the accuracy of the algorithm, i.e. the number of correctly classified instances of all assignments is 88.3%, indicating that the exploitation of pre-annotated morphosyntactic and lexical features can play a useful role in informing the annotation of complex sentences for internal IS. In particular, the algorithm assigned a relatively large number of false positives for the class NO INTERNAL IS in places where it is in fact present in the form of clause-internal contrastive focus, suggesting that contrast as an IS notion merits special attention with respect to annotation.

6. Conclusion

This paper responded to the challenge of annotating information structure in complex sentences by outlining certain desiderata with respect to both annotation format and the annotation process, informed by state-of-the-art theoretical knowledge, as well as practical issues identified for previous IS annotation schemes. In particular, the specific demands imposed by the IS properties of complex sentences were shown to add further weight to the importance of multidimensional, standoff annotation formats. With respect to the annotation process, a two-stage process was advocated for the IS annotation of dependent clauses (external IS, internal IS); for both stages, it was shown that rule-based algorithms which exploit pre-annotated non-IS features have the potential to play a useful role in the IS annotation of complex sentences in future.

7. Acknowledgements

This research was funded by the Research Foundation Flanders (FWO) via a postdoctoral fellowship awarded to the author [2021–2024, Grant no. 12ZL522N]. The author kindly thanks Anne Breitbarth for her valuable feedback on earlier versions of this paper, as well as two anonymous reviewers for their constructive comments.

8. Bibliographical References

Andresen, M., Vauth, M., and Zinsmeister, H. (2020). Modeling ambiguity with many annotators and self-assessments of annotator certainty. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 48–59, Barcelona, Spain, December. Association for Computational Linguistics.

Barteld, F., Ihden, S., Schröder, I., and Zinsmeister, H. (2014). Annotating descriptively incomplete language phenomena. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 99–104, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Baumann, S., Brinckmann, C., Hansen-Schirra, S., Kruijff, G.-J., Kruijff-Korbayová, I., Neumann, S., Steiner, E., Teich, E., and Uszkoreit, H. (2004). The MULI project: Annotation and analysis of information structure in German and English. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Beck, C., Booth, H., El-Assady, M., and Butt, M. (2020). Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain, December. Association for Computational Linguistics.

Bianchi, V. and Frascarelli, M. (2010). Is topic a root phenomenon? *Iberia: An International Journal of Theoretical Linguistics*, 2(1):43–88.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech communication*, 33(1-2):23–60.

Bohnet, B., Burga, A., and Wanner, L. (2013). Towards the annotation of Penn TreeBank with information structure. In Ruslan Mitkov et al., editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1250–1256, Nagoya. Asian Federation of Natural Language Processing.

Boyd, A. (2007). Discontinuity revisited: An improved conversion to context-free representations. In *Proceedings of the Linguistic Annotation Workshop*, pages 41–44, Prague, Czech Republic, June. Association for Computational Linguistics.

Boye, K. and Harder, P. (2007). Complement-taking predicates: usage and linguistic structure. *Studies in Language*, 31(3):569–606.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the workshop on Treebanks and Linguistic theories*, pages 24–41, Sozopol, Bulgaria.

Buráňová, E., Hajičová, E., and Sgall, P. (2000). Tagging of very large corpora: Topic-Focus articulation. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, pages 139–144, Saarbrücken, Germany, Universität des Saarlandes. Association for Computational Linguistics.

Büring, D. (2007). Semantics, intonation and information structure. In Gillian Ramchand et al., editors, *The Oxford handbook of linguistic interfaces*, pages 445–473. Oxford University Press, Oxford.

Bybee, J. (2002). Main clauses are innovative, subordinate clauses are conservative: consequences for the nature of constructions. In Joan Bybee et al., editors, *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson*, pages 1–17. John Benjamins, Amsterdam.

Cahill, A. and Riester, A. (2012). Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pages 232–236, USA. Association for Computational Linguistics.

Calhoun, S., Nissim, M., Steedman, M., and Brenier, J. (2005). A framework for annotating information structure in discourse. In Adam Meyers, editor, *Frontiers in Corpus Annotation II: Pie in the Sky, ACL2005 Conference Workshop*, pages 45–52, Ann Arbor, Michigan, June 2005.

Celano, G. G. A. (2019). Standoff annotation for the Ancient Greek and Latin Dependency Treebank. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brussels, Belgium*, pages 149–153, New York. Association for Computing Machinery.

Chafe, W. (1984). How people use adverbial clauses. In *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society*, pages 437–449, Berkeley, CA. Berkeley Linguistics Society.

Cook, P. and Bildhauer, F. (2011). Annotating information structure: the case of topic. In Stefanie Dipper et al., editors, *Beyond semantics: Corpus-based investigations of pragmatic and discourse phenomena*, Bochumer Linguistische Arbeitsberichte 3, pages 45–56. Ruhr-Universität Bochum, Sprachwissenschaftliches Institut, Bochum.

Cook, P. and Bildhauer, F. (2013). Identifying "aboutness topics": two annotation experiments. *Dialogue & Discourse*, 4(2):118–141.

Dahl, Ö. (1974). *Topic, comment, contextual boundedness and focus*. Buske, Hamburg.

Dalrymple, M. and Nikolaeva, I. (2011). *Objects and information structure*. Cambridge University Press, Cambridge.

De Cat, C. (2012). Towards an interface definition of root phenomena. In Lobke Aelbrecht, et al., editors, *Main Clause Phenomena: New Horizons*, pages 135–158. John Benjamins, Amsterdam.

de Swart, H. and de Hoop, H. (2014). Topic and focus. In Lisa Cheng et al., editors, *The First Glot International State-of-the-Article Book: The Latest in Linguistics*, pages 105–130. de Gruyter, Berlin.

Dehé, N. and Wichmann, A. (2010). Sentence-initial *I think (that)* and *I believe (that)*: prosodic evidence for uses as main clause, comment clause and discourse marker. *Studies in Language*, 34(1):36–74.

Depraetere, I. (1996). Foregrounding in English relative clauses. *Linguistics*, 34:699–731.

Diessel, H. (2001). The ordering distribution of main and adverbial clauses: A typological study. *Language*, 77:433–455.

Dipper, S. (2005). Xml-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of the Berliner XML Tage 2005 (BXML 2005)*, pages 39–50. Berlin, Germany.

Ebert, C., Ebert, C., and Hinterwimmer, S. (2014). A unified analysis of conditionals as topics. *Linguistics and Philosophy*, 37(5):353–408.

Erteschik-Shir, N. (2007). *Information structure: the syntax-discourse interface*. Oxford University Press, Oxford.

Fabb, N. (1990). The difference between English restrictive and nonrestrictive relative clauses. *Journal of Linguistics*, 26(1):57–77.

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S., and Stoel, R. (2007). Information structure. In Steffi Dipper, et al., editors, *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, pages 147–187. Universitätsverlag Potsdam, Potsdam.

Gussenhoven, C. (1999). On the limits of focus projection in English. In Peter Bosch et al., editors, *Focus: linguistic, cognitive, and computational perspectives*, pages 43–55. Cambridge University Press, Cambridge.

Haegeman, L. (2007). Operator movement and topicalisation in adverbial clauses. *Folia Linguistica*, 41(3/4):279–325.

Haiman, J. (1978). Conditionals are topics. *Language*, 54(3):564–589.

Hajicová, E., Partee, B. B. H., and Sgall, P. (1998). *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer, Dordrecht.

Hedberg, N. (2013). Multiple focus and cleft sentences. In Katharina Hartmann et al., editors, *Cleft structures*, pages 227–250. John Benjamins, Amsterdam.

Hempelmann, C. F., Dufty, D., McCarthy, P. M., Graesser, A. C., Cai, Z., and McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In Bruno Bara, et al., editors, *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, pages 941–946, Mahwah, NJ. Erlbaum.

Hooper, J. B. and Thompson, S. A. (1973). On the applicability of root transformations. *Linguistic Inquiry*, 4(4):465–497.

Hooper, J. B. (1975). On assertive predicates. In John P. Kimball, editor, *Syntax and Semantics*, volume 4, pages 91–124. Academic Press, San Diego, CA.

Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3-4):211–225.

Kiparsky, P. and Kiparsky, C. (1970). Fact. In M Bierwisch et al., editors, *Progress in linguistics*, pages 143–147. Mouton, The Hague.

Koktová, E. (1996). Wh-extraction and the topic-focus articulation of the sentence. In Barbara H. Partee et al., editors, *Discourse and Meaning: Papers in Honor of Eva Hajičová*, pages 255–271. John Benjamins, Amsterdam.

Komagata, N. (2003). Information structure in subordinate and subordinate-like clauses. *Journal of Logic, Language and Information*, 12(3):301–318.

Krifka, M. and Musan, R. (2012). Information structure: overview and linguistic issues. In Manfred Krifka et al., editors, *The expression of information structure*, pages 1–44. de Gruyter, Berlin.

Krifka, M. (2007). Basic notions of information structure. In Caroline Féry et al., editors, *Interdisciplinary Studies on Information Structure*, pages 13–56. Universitätsverlag, Potsdam.

Lahousse, K. and Borremans, M. (2014). The distribution of functional-pragmatic types of clefts in adverbial clauses. *Linguistics*, 52(3):793–836.

Lahousse, K. (2010). Information structure and epistemic modality in adverbial clauses in French. *Studies in Language*, 34(2):298–326.

Lahousse, K. (2022). Is focus a root phenomenon? In Davide Garassino et al., editors, *When data challenges theory: unexpected and paradoxical evidence in information structure*, pages 148–182. John Benjamins, Amsterdam.

Laprun, C., Fiscus, J., Garofolo, J., and Pajot, S. (2002). Recent improvements to the ATLAS architecture. In *Proceedings of the Second International Conference on Human Language Technology (HLT'02)*, pages 263–268.

Leech, G. (2005). Adding linguistic annotation. In Martin Wynne, editor, *Developing linguistic corpora: a guide to good practice*, pages 17–29. Oxbow Books, Oxford.

Lehmann, C. (1984). *Der Relativsatz: Typologie seiner Strukturen, Theorie seiner Funktionen, Kompendium seiner Grammatik*. John Benjamins, Amsterdam.

Lehmann, C. (1988). Towards a typology of clause linkage. In John Haiman et al., editors, *Clause combining in grammar and discourse*, pages 181–225. John Benjamins, Amsterdam.

Lelandais, M. and Ferré, G. (2017). How are three syntactic types of subordinate clauses different in terms of informational weight? *Anglophonia*, 23. `http://journals.openedition.org/anglophonia/1200`.

Lüdeling, A., Ritz, J., Stede, M., and Zeldes, A. (2016). Corpus linguistics and information structure research. In Caroline Féry et al., editors, *The Oxford handbook of information structure*, pages 599–620. Oxford University Press, Oxford.

Maier, W. and Lichte, T. (2011). Characterizing discontinuity in constituent treebanks. In Philippe de Groote, et al., editors, *Formal Grammar*, pages 167–182, Berlin, Heidelberg. Springer.

Maki, H., Kaiser, L., and Ochi, M. (1999). Embedded topicalization in English and Japanese. *Lingua*, 1(109):1–14.

Marchese, L. (1977). Subordinate clauses as topics in Godie. In Matin Mould et al., editors, *Papers from the 8th Conference on African Linguistics*, pages 157–164, Los Angeles, CA. Department of Linguistics, University of California.

Markert, K., Hou, Y., and Strube, M. (2012). Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, page 795–804, USA. Association for Computational Linguistics.

Mathesius, V. (1975). *A functional analysis of present day English on a general linguistic basis*. Mouton, The Hague.

Matić, D., Van Gijn, R., and Van Valin Jr, R. D. (2014). Information structure and reference tracking in complex sentences. In Rik van Gijn, et al., editors, *Information structure and reference tracking in complex sentences*, pages 1–42. John Benjamins, Amsterdam.

Matsuda, K. (1998). On the conservatism of embedded clauses. In Monika S. Schmid, et al., editors, *Historical Linguistics 1997: Selected papers from the 13th International Conference on Historical Linguistics, Düsseldorf, 10–17 August 1997*, pages 255–268. John Benjamins, Amsterdam.

Merten, M.-L. and Seemann, N. (2018). Analyzing constructional change: Linguistic annotation and sources of uncertainty. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18, page 819–825, New York, NY, USA. Association for Computing Machinery.

Milićev, T. and Milićević, N. (2012). Leftward movement with discontinuous appositive constructions. *Acta Linguistica Hungarica*, 59(1-2):205–220.

Neeleman, A., Titov, E., Van de Koot, H., and Vermeulen, R. (2009). A syntactic typology of topic, focus and contrast. In Jeroen van Craenenbroeck, editor, *Alternatives to cartography*, pages 15–52. de Gruyter, Berlin.

Nikolaeva, I. (2001). Secondary topic as a relation in information structure. *Linguistics*, 39(1):1–49.

Nissim, M. (2006). Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, page 94–102, USA. Association for Computational Linguistics.

Paggio, P. (2006). Annotating information structure in a corpus of spoken Danish. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC2006)*, pages 1606–1609, Genova, Italy.

Partee, B. H. (1996). Allegation and local accommodation. In Barbara H. Partee et al., editors, *Discourse and meaning: papers in honor of Eva Hajicová*, pages 65–86. John Benjamins, Amsterdam.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman, London.

Rahman, A. and Ng, V. (2012). Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, page 798–807, USA. Association for Computational Linguistics.

Reinhart, T. (1981). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1):53–94.

Riester, A., Brunetti, L., and De Kuthy, K. (2018). Annotation guidelines for Questions under Discussion and information structure. In Evangelia Adamou, et al., editors, *Information structure in lesser-described languages: studies in prosody and syntax*, pages 403–443. John Benjamins, Amsterdam.

Riester, A. (2009). Stress test for relative clauses. In Arndt Riester et al., editors, *Focus at the syntax-semantics interface*, pages 69–86. University of Stuttgart, Stuttgart.

Romary, L., Zeldes, A., and Zipser, F. (2010). [Tiger2/] documentation [Technical Report]. inria-00593903v2.

Schachter, P. (1973). Focus and relativization. *Language*, 41(1):19–46.

Schiffrin, D. (1992). Conditionals as topics in discourse. *Linguistics*, 30:165–197.

Schilder, F. and Tenbrink, T. (2002). The interplay of information structure and the placement of *after* and *before*. In *Proceedings of the Workshop on Information Structure in Context*, Stuttgart. University of Stuttgart.

Smith, N., Hoffmann, S., and Rayson, P. (2008). Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and Linguistic Computing*, 23(2):163–180.

Song, S. (2014). Information structure of relative clauses in English: a flexible and computationally tractable model. *Language and Information*, 18(2):1–29.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5–6):701–721.

Stede, M. and Mamprin, S. (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, May. European Language Resources Association (ELRA).

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689.

Surányi, B. (2007). Focus structure and the interpretation of multiple questions. In Kerstin Schwabe et al., editors, *On information structure, meaning and form*, pages 229–253. John Benjamins, Amsterdam.

Taylor, A., Marcus, M., and Santorini, B. (2003). The Penn Treebank: an overview. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*, pages 5–22. Kluwer, Dordrecht.

Thompson, S. A. and Longacre, R. E. (1985). Adverbial clauses. In Timothy Shopen, editor, *Language typology and syntactic description*, volume 2, pages 171–234. Cambridge University Press, Cambridge.

Thompson, S. A. (1985). Grammar and written discourse: Initial vs. final purpose clauses in English. *Text-Interdisciplinary Journal for the Study of Discourse*, 5(1-2):55–84.

Tophinke, D. (2012). Syntaktischer Ausbau im Mittelniederdeutschen: Theoretisch-methodische Überlegungen und kursorische Analysen. *Niederdeutsches Wort*, 52:19–46.

Umbach, C. (2006). Non-restrictive modification and backgrounding. In *Proceedings of the Ninth Symposium on Logic and Language*, pages 152–159, Budapest. Hungarian Academy of Sciences.

Vallduví, E. and Zacharski, R. (1994). Accenting phenomena, association with focus, and the recursiveness of focus-ground. In P. Dekker et al., editors, *Proceedings of the 9th Amsterdam Colloquium*, pages 683–702. ILLC (Institute for Logic, Language and Computation)/Department of Philosophy, Amsterdam.

Vallduví, E. (1992). *The informational component*. Garland Press, New York.

van Kuppevelt, J. (1995). Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1):109–147.

von Fintel, K. (2004). Would you believe it? The King of France is back! (presuppositions and truth-value intuitions). In Marga Reimer et al., editors, *Descriptions and beyond*, pages 315–341. Oxford University Press, Oxford.

von Heusinger, K. (1999). Intonation and information structure. University of Konstanz. Habilitationsschrift.

von Stutterheim, C. and Klein, W. (1989). Referential movement in descriptive and narrative discourse. In Rainer Dietrich et al., editors, *Language processing in social context*, pages 39–76. North-Holland, Amsterdam.

Wiklund, A.-L., Bentzen, K., Hrafnbjargarson, G. H., and Hróarsdóttir, Þ. (2009). On the distribution and illocution of V2 in Scandinavian *that*-clauses. *Lingua*, 119(12):1914–1938.

Ziai, R. and Meurers, D. (2018). Automatic focus annotation: Bringing formal pragmatics alive in analyzing the information structure of authentic data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 117–128, New Orleans, Louisiana, June. Association for Computational Linguistics.

9. Language Resource References

Beck, J., Ecay, A., and Ingason, A. K. (2015). Annotald. version 1.3. 7.

Booth, H., Breitbarth, A., Ecay, A., and Farasyn, M. (2020). A Penn-style treebank of Middle Low German. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 766–775, Marseille, France, May. European Language Resources Association.

Druskat, S., Gast, V., Krause, T., and Zipser, F. (2016). corpus-tools.org: An interoperable generic software tool set for multi-layer linguistic corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4492–4499.