Why Jesus and Job spoke bad Welsh : the origin and distribution of V2 orders in Middle Welsh

Meelen, M.

CHAPTER 3

Coding features relevant for Information Structure

If we want to determine to what extent - if at all - Information Structure (IS) relates to word order (change), we first need an adequate description of IS and its relevant notions in the grammar of historical Welsh. Although IS is a relatively new subfield of pragmatics (cf. Meurman-Solin, López-Couso, and Los (2012:3)), there is a vast literature on IS-related phenomena in a great number of languages. A general consensus on the exact definition of most information-structural notions expressed in the grammar is, however, still lacking.

Apart from defining information structure and its place in linguistic research, this chapter aims to provide an overview of those interpretive notions that are considered to be information-structural primitives. The grammar of a language has several means at its disposal to express information structure, but only those relevant to the present diachronic research will be discussed in detail.

Although recent overviews by Krifka (2008), Ritz, Dipper, and Götze (2008), Traugott and Pintzuk (2008) and, in particular, Götze et al. (2007) are insightful, there is no generally accepted or standardised way of coding IS features systematically yet. In this chapter I argue that any good description of the information structure of a language at the very least contains a detailed overview of how the grammar of the language expresses the core notions of **givenness**, **topic-comment** and **focus-background** (cf. section 3.3). I furthermore provide step-by-step guidelines on the procedures of coding those IS features. I conclude this chapter with a methodological note on the strategies implemented in the rest of this thesis to find the right mappings of information-structural primitives to the expressed word order types.

3.1 What is Information Structure?

> *"Terminological profusion and confusion, and underlying conceptual vagueness, plague the relevant literature to a point where little may be salvageable (...) In addition there is reason to think that the whole area may be reducible to a number of different factors (...)."*
>
> (Levinson, 1983:x)

The whole field of information structure (or, in fact, 'confusing' terminology like 'topic/comment' or 'theme/rheme') belongs to a long list of topics Stephen Levinson chooses *not* to discuss in his textbook on pragmatics. Some ten years later, Knud Lambrecht proposes a new theory of sentence formation, because there "still is disagreement and confusion" about information structure, a term he borrows from Halliday (1967) for a "grammatical component" of language. Another decade passes and Kruijff and Duchier (2003) are still concerned with the 'proliferating terminologies', to the extent that they find it necessary to add an insightful diagram to their paper visualising the 'terminological profusion and confusion' that seems to have haunted the field since the 1980s.

The *profusion* is indeed partly responsible for the enduring *confusion*. Using two (or three or even more) terms for one and the same phenomenon is often misleading. Employing just one of those terms to describe different phenomena at the same time is downright ambiguous. From that perspective, Vallduví and Vilkuna's *kontrast* with a *k*, no matter how well-argued for, perfectly illustrates the field's confused history (cf. Vallduví and Vilkuna (1998)).

Difficulty in *defining* information structure other than 'a subfield' (of pragmatics or semantics) contributed to the afore-mentioned confusion as well. Most collections of papers describing IS phenomena in various languages that bother to give a definition, resort to explaining what IS *does* or what it is *not*, rather than what it *is*. Examples of those information-structural effects include "encoding of the relative salience of the constituents of a clause" (Foley, 1994:1678), "presentation of information as old and new" (De Swart & De Hoop, 1995:3) and "packaging of information" (cf. Féry and Krifka (2008:2) following Chafe (1976)). Other common 'definitions' actually aim to identify the place of IS in relation to various linguistic notions, cognitive domains or as an in-between 'interface issue' (Mereu, 2009:2).

This brief introduction does not solve any issues in information-structural theory, it merely serves to illustrate the difficulty in choosing the right terminology on the one hand, and the necessity to give a detailed overview of the methodological considerations on the other. I use Zimmermann & Féry's definition of IS mediating "between the modules of linguistic competence in the narrow sense, such as syntax, phonology, and morphology, and other cognitive faculties which serve the central purpose of the fixation of belief by way of information update, pragmatic reasoning, and general inference processes." (Zimmermann & Féry, 2010:1). This notion is fully compatible with the Communicative model of Common Ground, which I use as a starting point for the present overview of the IS annotation guidelines (see section 3.2).

3.1.1 Brief history of IS research

The systematic study of the pragmatic organisation of discourse has its origin in the theory of the 'Functional Sentence Perspective' by the Prague Linguistic Circle initiated by Vilém Mathesius (1882-1945) (cf. Nekula (1999) and Mereu (2009)). His work on functional linguistics (Mathesius, 1929 [1983]) showed that the presentation of given material (the theme) and new material (the rheme) plays an important role in the structure of a language. Later scholars of the Prague School like Firbas (1964) employed the gradient notion of Communicative Dynamism (CD) to account for information structural phenomena, arguing that CD is responsible for the linear arrangement of syntactic constituents. Elements in the sentence with 'least CD' (i.e. the theme or topic or that which is contextually known) precede those with 'more CD' (i.e. those conveying new or unlinked information) (cf. Erteschik-Shir (2007:2)).

The notion of Common Ground (CG) was introduced by Paul Grice in the William James lectures of 1966-1967 as a term for the presumed background information or 'the context' of a conversation (cf. Stalnaker (1974), Grice (1989), Stalnaker (2002) and 3.2 below). Chafe (1976) first discussed semantic distinctions used in 'information packaging' (adopted in a formal context by Vallduví (1992)). Typological research in the late 1970s and 1980s by Li and Thompson (1976) and Mithun (1987) distinguished subject- and topic-oriented, or syntactically- or pragmatically-based languages. Givón (1984:204) argued that word order variation is "controlled by discourse-pragmatic considerations pertaining to new vs. old, topical vs. non-topical, discontinuous vs. disruptive information".

Following this, various researchers in the late 1980s and 1990s investigated focus structures (Abraham & de Meij, 1986) or topic structures (cf. Reinhart (1982), Lambrecht (1994), É.Kiss (1995), Dik (1997) and Büring (1997)) or a hierarchy of both topic and focus, see Payne (1987), Choi (1999), Frascarelli (2000) and Mereu (2009) for an overview).

In 2003, researchers from the universities of Potsdam and Berlin founded the 'Collaborative Research Center (Sonderforschungsbereich / SFB 632)' on Information Structure. Between 2003 and 2015, a grand total of 19 projects and 53 researchers aimed to formulate integrative models of information structure in various disciplines of linguistics and human cognition. They defined information structure as 'the structuring of linguistic information, typically in order to optimise information transfer within discourse.' (see the project description on their website `www.sfb632.uni-potsdam.de`). Research output of this centre focusses on the interaction of the relevant formal linguistic levels, general cognitive processing of information structure and finally on a cross-linguistic typology of information structural devices.

In an attempt to provide an insightful overview of what has by now become a (linguistic) field of its own, the *Handbook of Information Structure* will be published by Oxford University Press in the course of 2016 (Féry & Ishihara, 2016).

3.1.2 Where is information structure?

Information structure is usually mentioned as a subfield of pragmatics within the field of linguistics (cf. Meurman-Solin et al. (2012:3)), because it is related to language use: the relation of signs to those who interpret the signs. According to Kruijff and Duchier (2003:249), both utterance-internal (IS) as well as utterance-external semantic devices interact to provide the discourse context. IS is thus closely related to discourse analysis and semantics.

The question 'Where is information structure?' *in language*, rather than *in the field of linguistics* is far more interesting, but also more difficult to answer. Is it a 'grammatical component' as Lambrecht (1994:xiii) suggested? Is it part of (or encoded in) syntax, semantics or phonology? Or do IS phenomena operate on the interfaces of all of those (cf. Mereu (2009:2))?

Functional theories of language focus on what information structure contributes to the grammar (cf. Kuno (1987) and Dik (1997)). In a similar way, Role and Reference grammar, as employed by, among others, Van Valin (1993b), stores grammatical structures as constructional templates with specific sets of morphosyntactic, semantic and pragmatic properties, so that they are naturally linked (cf. Erteschik-Shir (2007:4-5)). Jackendoff (1972) and Horvath (1981) formalised discourse-semantic notions in structural relations, paving the way for discourse-configurational approaches (e.g. É.Kiss (2001)) in which topic and focus are linked to particular structural positions and thus part of the syntax. Further within Generative Grammar then, in particular in Rizzi's Cartography (cf. Rizzi (1997) and Rizzi (2004)), information structural features surface as separate projections in the sentence peripheries.

However, if information structure plays a role in semantics and phonology as well as in syntax, these representations make it difficult to express IS notions in a unified and systematic way. Alternatives to cartographic approaches by, among others, Neeleman and Van de Koot (2008) and Kučerová and Neeleman (2012) aim to solve this by mapping the syntax to the information structure at the interfaces. Multi-layered theories like lexical-functional grammar (LFG), head-driven phrase structure grammar (HPSG) or combinatory categorial grammar (CCG) take a different approach by formalising information structure in a way equal to the status of the other components of grammar (cf. Erteschik-Shir (2007:4)).

3.1.3 Main questions in IS research

As is clear from the above introduction, there are still many questions in information-structural research left unanswered. Even the exact object or unit of investigation varies from study to study. It is clear that IS phenomena can be observed by studying sentences in their context, but is that the only way? Can certain IS-related expressions also occur on the sentence or clause level, or possibly even on lower ranks of syntactic structure (cf. Kruijff and Duchier (2003:251))?. Information structure seems multi-modular and multi-levelled: an exhaustive investigation of IS phenomena in a language thus requires input from various aspects of the grammar

(syntax, semantics, morphology and phonology), but also from interacting cognitive domains (pragmatic reasoning, the fixation of belief and the update of information states, (cf. Zimmermann and Féry (2010:2)). In this chapter, I relate all coded IS notions to their grammatical markings as well as the way they function in our brain.

What are the basic notions or dimensions of IS?

Information-structural phenomena can be found in various parts of the grammar of a language, but what is it exactly that we are trying to find? The 'profusion' of terminology mentioned in the introduction hardly makes it easier to define the basic notions of IS. Recent IS literature, however, has not only described certain phenomena in a particular language, but also aimed to find the core dimensions or primitives of information structure. Kruijff and Duchier (2003:251) identify two recurrent patterns: "topic/comment" or "theme/rheme" and "background/kontrast" or "given/new". Zimmermann and Féry (2010:1) separate the second notion and claim that there are three basic concepts of IS:

— focus vs. background
— topic vs. comment
— given vs. new

Kučerová and Neeleman (2012:1) agree stating "these notions may require refinements and subdivisions, but there does not seem to be a substantial case in the literature for extending the set.". In other words, there seem to be no languages that, for example, have a separate class for elements that are neither new nor given with a specific syntactic distribution.

There is one important notion of IS, however, that has not been mentioned so far, namely 'contrast' (or 'kontrast', following Vallduví and Vilkuna (1998)). Intuitively, contrast is associated with an element of rejection or correction. Contrastive focus often emphasises one particular alternative. Repp (2010:1338) points out, however, that "contrast does not necessarily involve an element of rejection". In an earlier paper, Krifka (1999) already pointed out that contrastive focus can also be additive and furthermore, that contrast does not have to be associated with focus structures, because contrastive topics can also be found (cf. Krifka (2008)).

I therefore do not treat contrast as an IS primitive, but rather discuss the contrastive examples as they occur in one of the above-mentioned dimensions. These three dimensions will form the basis of my methodological analysis and IS annotation scheme.

How can IS be expressed in the grammar?

Knowing what to look for is one thing, knowing what it looks *like* in a language is a very different question. The great number of publications on IS phenomena is partly due to the many ways in which IS can be expressed. Examples can be found in a wide variety of languages in one or more of the following grammatical components:

- **Phonology**, in particular prosodic devices like pitch accent, deaccenting, and, as an extreme form of deaccenting, complete phonological reduction or ellipsis. Intonational phrases can also be used to indicate topics in English, German or Japanese (Krifka & Musan, 2012:34).

- **Morphology**. Some languages have special suffixes to mark, for example, VP focus, such as the perfective *-go* on the verb in Chadic (cf. Hartmann and Zimmermann (2007)) or the *no/gon* morphemes in Tsez (cf. Kučerová and Neeleman (2012:2)).

- **Syntax** can express IS phenomena in different ways: particular positions or word order patterns (e.g. fronting), agreement or the lack thereof (e.g. in a language like Tsez, cf. Kučerová and Neeleman (2012) or Middle Welsh, see Chapter 5) and specific constructions, such as cleft or pseudo-cleft sentences that are well-known in English.

- **Lexical items** related to certain IS phenomena come in various kinds: specific topic or focus particles, adverbials or determiners or anaphoric expressions.

What are the mapping rules between IS dimensions and expressions?

There are still many questions about the exact relation between information structure and the above-mentioned components of grammar. Some generalisations can be clearly formulated when it comes to IS and phonology: there seem to be no languages, for example, in which "old material must be stressed and new material de-stressed", which, according to Kučerová and Neeleman (2012:19), can hardly be a coincidence. The extent to which, and how exactly, IS is integrated in syntax and semantics is still an open question too, although "[T]here appears to be general agreement in the field that it would be more desirable for information structure and semantics to be part of the same system" (Kučerová & Neeleman, 2012:18).

The present thesis is concerned with the interaction of information structure and word order change. Therefore, although some elements of other grammatical components are coded, the syntactic way(s) of expressing IS in Welsh will be the main focus of my analysis. How this is implemented exactly will be discussed in Chapter 5.

3.1.4 Why study Information Structure?

Information structure is an integral part of human language, making the study of it invaluable in any effort to fully understand and describe the grammar and underlying mechanisms of a language. IS research can in particular shed light on variation and 'free' alternations found in languages, such as OV/VO word order, particle verbs (*He carried out the instructions. vs. He carried the instructions out.*) and the well-known dative alternation (*He give Sarah the book. vs. He gave the book to Sarah.*). Upon closer look at their information-structural status, these subtle

alternations very often turn out to be less 'free' than previously thought. Better insight in IS mechanisms can therefore be useful in the more applied field of L2 acquisition, designing textbooks and grammars that help learners to gain the much-desired native-speaker fluency (cf. Hannay and Mackenzie (2002) and Lozano (2006)).

But variation is also frequently encountered (and not always sufficiently explained) in diachronic data. Again, studying the information-structural properties of the specific alternations might shed more light on why changes in the language occurred, and, even more interestingly, why changes developed in one way and not the other. The information-structure background of the change in Welsh word order therefore serves as an excellent example.

3.1.5 Information Structure in diachronic data

Studying IS in diachronic data also has its limitations. Most of these have their origin in limited access to the data, which in turn, is only available in a limited form, i.e. only written sources survived. An additional problem for at least some of these sources is that we cannot always be sure to what extent they represent the language as it was used in a particular time or place (if, in fact, we know when and where that was in the first place) (cf. Meurman-Solin et al. (2012:10)). Is the manuscript version that survived merely a rendition of a story that clearly belonged in an oral tradition? If so, to what extent was it reworked - if at all - to fit the written medium? There is a clear stylistic difference between written and spoken language, so how can we evaluate any variation we encounter if we are not sure to which broad genre the text belongs in the first place? In general, the lack of information that may convey crucial IS differences such as intonation, is problematic. If prosody played an important role in marking IS patterns in the language, its impact is difficult to ascertain (although some research on prosodic phrases and stress patterns in historical data has been carried out (cf. Speyer (2008) and Hinterhölzl (2009)). Finally, the lack of native speaker judgments or possibility to run psycholinguistic experiments means traditional tests for specific IS patterns cannot be carried out. Certain particles or questions testing the scope of focus constructions, for example, such as *What happened?* or *Who did you see?* are simply not always available in the data (Traugott & Pintzuk, 2008:63).

We thus have to work with the data we have, limited as it may be, and a certain amount of caution is necessary in drawing far-reaching conclusions from results based on data with an uncertain philological background. As long as we are aware of what the data *can* tell us, studies of IS in diachronic data form an invaluable contribution to the description of older stages of the language and how it developed. Starting from the Common Ground, the rest of this chapter provides an overview of the most important notions of IS discussed above to describe the annotation scheme used for the historical Welsh database. The IS notions are discussed in relation to the two important elements of Zimmermann and Féry's (2010) definition of IS: their cognitive reality and the way they can be expressed or marked in the grammar.

3.2 Information Packaging & Common Ground

> *"Once upon a time there was a man who went to cut firewood in the forest above his village in the depths of winter. As he was cutting branches from a tree on the edge of a cliff he missed his footing and fell into the gorge, and resigned himself to a certain death on the rocks below. As it happened, there was a hibernating dragon in the gorge, and it opened its jaws in a great yawn just in time to catch the falling woodcutter."*
>
> (Ramble, 2013:75)

Storytelling, like any other act of discourse (reading a book, talking to a friend, listening to the radio, etc.[1]), involves the transfer of information. Successful communication of coherent discourse (making the reader/listener understand) depends at least partly on the optimisation of this information transfer, relative to the temporary needs of interlocutors (cf. Krifka (2008:15)).[2]

Stalnaker (1974) and Karttunen (1974) used the Gricean concept of Common Ground (CG) "as a way to model the information that is mutually known to be shared and continuously modified in communication" Krifka (2008:15). According to Krifka, the CG contains both a set of mutually accepted **propositions** as well as a set of **entities** that have been introduced into the CG before. As the discourse develops, the CG changes continuously and therefore the information has to be 'packaged in correspondence with the CG at the point at which it is uttered' (cf. Krifka (2008:16) following Chafe (1976)'s "Information Packaging"). As Stalnaker (2002) points out, the Common Ground is not necessarily the same as our Common Belief, i.e. the presuppositions of speakers, listeners, readers and writers. The Common Ground defines the context only, irrespective of whether the propositions uttered in a particular context are true or believed to be true.

There can be a divergence between the assumed context or Common Ground and people's actual beliefs. This is seen in Von Fintel's example of a daughter informing her father she is getting married with the words: "O Dad, I forgot to tell you that my fiancé and I are moving to Seattle next week" (Von Fintel, 2000:9). Even though the proposition about the engagement is new to her father, her daughter has decided to present it as old news in the context, because, for example, she does not want to discuss it further. Her father can then choose to grant his daughter's wish by accepting this context along with its subtext (i.e. she does not want to talk about it), even though their initial common beliefs about the daughter's relationship status were very different. Stalnaker (2002:716) therefore points out

[1] Note that I use the term "discourse act" in the sense of any piece of communication, both oral and written (cf. Di Eugenio (2003)). This linguistic interpretation does not include the Foucauldian sense of 'discourses of knowledge', which usually does not involve any textual analysis (cf. Fairclough (1992) and Bucholtz (2008)). Its use here is broader than just 'Conversation Analysis' in sociocultural linguistics.

[2] In spoken direct discourse like conversations, optimal communication is based on the cooperative principle of the four Gricean maxims of quantity, quality, relation and manner (Grice, 1989). Since the current study investigates historical data, I only focus on written texts in the rest of this discussion on discourse structure.

that the common ground "should be defined in terms of a notion of *acceptance* that is broader than the notion of belief". In the following section, I turn to how this kind of model of the Common Ground relates to text comprehension in our brain, this concerns the *accepted* context, irrespective of whether this corresponds to the parties' actual common beliefs.

3.2.1 Text comprehension in our brain

How do we interpret any form of discourse in the first place? The main reason we can understand the opening paragraph of the (originally Tibetan) woodcutter's tale cited in the beginning of this section is because we know the meaning of the individual words and because of the coherence between the sentences. Coherence between sentences (the systematically structured passages of discourse) is one "of the most fundamental characteristics of texts" (Schmalhofer, Friese, Pietruska, Raabe, & Rutschmann, 2005:1949). There are various ways in which textual coherence can be established, e.g. Schmalhofer et al. (2005:1949):

(1) a. anaphora resolution (cf. Glenberg, Meyer, and Lindem (1987))
    b. identifying overlaps in arguments of different propositions (cf. Kintsch and Van Dijk (1978))
    c. memory processes resonating for words with closely related meanings (cf. O'Brien, Rizzella, Albrecht, and Halleran (1998))
    d. inference processes driven by a search for meaning (cf. Graesser, Singer, and Trabasso (1994))

Psycho- and neurolinguistic experiments can provide insights on how our brain works when we are reading a text. Brain imaging techniques such as electroencephalography (EEG) measuring electrical activity of brain waves and functional magnetic resonance imaging (fMRI) can be used to shed more light on the processes mentioned in (1) (cf. Ferstl and von Cramon (2001) and Hagoort, Hald, Bastiaansen, and Petersson (2004)). Event-related potentials or 'ERP effects', in particular, are useful in linguistic research, because they are the results of the electrical activity of brain waves in relation to the event of interest (a word/sentence/construction etc) measured by EEG (cf. Luck (2005) and Sprouse and Lau (2013)). Negative and positive peaks in this EEG activity can indicate mismatches in particular linguistic domains. A problem in anaphora resolution, for example, yields a sustained negative offset after 300ms: the 'Nref effect' (cf. Van Berkum, Koornneef, Otten, and Nieuwland (2007:160) and Komen (2013:27)). To illustrate this, consider the first two sentences of the woodcutter's tale again in (2):

(2) a. *Once upon a time there was a man who went to cut firewood in the forest above his village in the depths of winter.*
    b. *As he was cutting branches from a tree on the edge of a cliff*
    c. *he missed his footing and fell into the gorge,*
    d. *and resigned himself to a certain death on the rocks below.*

When reading a sentence like (2a), we hold as much information as possible in our working memory. However, instead of trying to store the separate words we read, we try to extract the ideas they represent (cf. Kintsch (1989)). Following Komen (2013:28), I call the representational form of the linguistic expression we build in our mind a 'mental entity'. The syntactic phrase *a man who...* is the linguistic expression that first of all refers to this created mental entity. The mental entity in its turn refers to "real-world concepts or to imaginary ones" (Komen, 2013:28) or its denotation (cf. Krifka (2008)) (in this case a man who is cutting firewood). Zwaan and Radvansky (1998) show that we dynamically transform every part of the discourse into a "situation model" consisting of a set of participants (the mental entities) and a set of propositions (actions or relationships involving these mental entities) (cf. Van Dijk and Kintsch (1983) and Kintsch and Rawson (2005) on propositional representations in the situation model, or the similar "mental model" as it is called by Craik (1943), Johnson-Laird (2013)). Figure 3.1 shows a schematic representation of Mental Entities in the Situational Model applied to our woodcutter's tale.

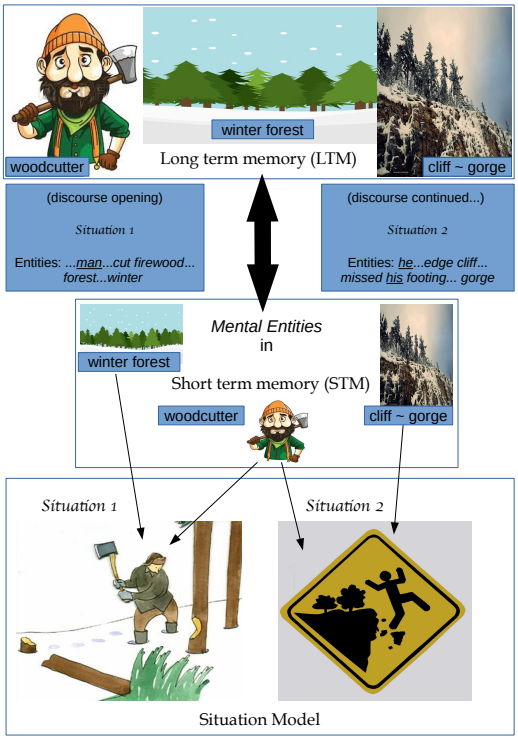


Figure 3.1: Schematic representation of entities in Long and Short-term memory

When we continue to read (2b), we dynamically update the model we built in our working memory describing the situation in which a particular mental entity, *a man*, is involved with certain propositions: he is cutting wood in a forest, the forest is above his village, it is the middle of winter, etc. As we parse (2b), we create a new mental entity in our working memory of the first linguistic expression we encounter, the pronoun *he*. Since pronouns are anaphors, we start a process of reference resolution (process (1a) above) in which we try to determine whether this mental entity matches with an already existing mental entity in the "situation model" (cf. Komen (2013:30)). In this case there is a perfect match with the mental entity we created to refer to *a man* in the previous sentence, so the features/characteristics of the phrase are added to the existing entity. Note that if (2b) were to have continued with *As she...*, we would have encountered a mismatch in gender (in English, *a man* cannot be referred to as *she*) resulting in the above-mentioned Nref effect in an experimental setting (as shown in various contexts by, among others, Van Berkum et al. (2007:160)). When we continue reading we further update our model with the propositions concerning the fact that the man is now cutting branches from a tree and that this tree is on the edge of a cliff, etc. Since the story goes on to relate how the same man *who went out to cut firewood* is now, in fact, *cutting branches from a tree*, there is a clear overlap in the arguments (see processes (1b) and (1c) above). The *edge of a cliff* in (2b) and the *gorge* in (2d) are another good example of this overlap in meaning. When parsing the rest of the sentence, we continue updating our model by adding and matching new mental entities and propositions. These propositions are not necessarily all found in the text itself: we can also access propositions that are stored in our long-term memory. We may for example associate *the depths of winter* in (2a) with a lot of snow, which in turn may result in a dangerous situation when you are busy working *on the edge of a cliff*.[3] We fully understand the following dramatic events in (2c), because we could make the right inferences (see process (1d) above) from the preceding context (i.e. working on the edge of a cliff in winter may be dangerous). We have just created a situation model in which the woodcutter is headed for a certain death, because he is falling into the rocky gorge. But now we continue to read this:

(3) *As it happened, there was a hibernating dragon in the gorge, and it opened its jaws in a great yawn just in time to catch the falling woodcutter.*

The scenario in which the man does *not* die was not part of our situation model: we did not expect this to happen especially not after the man himself pictured his 'certain death'. The developments in (3) are new and unexpected and we will have to create a new situation model containing the possibility of the man surviving the fall, or, at the very least, of the man not dying because he hit rock bottom, but because he was eaten by a dragon. According to Johnson-Laird (1989), it is easier to comprehend passages that lead unambiguously to a single model than

[3] For the potential audience of this particular tale, the inhabitants of the Tibetan plateau, this association will be even more accessible than for those living in much warmer areas of the world, but this only proves the point of 'optimal communication' in discourse.

passages that lead to multiple models. Again, we see that we do not just rely on the text to find the meaning of the passages, we also incorporate it in a broader context containing our knowledge of the physical, social and cultural world in which the discourse is presented. Bearing this in mind, the passage in (3) might be more accessible for the potential audience, the inhabitants of the Tibetan plateau (where dragons feature in many stories). Since we can only hold a limited number of models in our working memory at any given time (Johnson-Laird, Byrne, & Schaeken, 1992), we will soon discard the incorrect models to make room for new ones. With the next passage in the woodcutter's tale, we can finally reject the scenario involving the man's certain death:

(4) *The man survived the winter in the warmth of the sleeping dragon's maw, sustaining himself on the edible jewels that lay about the place in abundance.*

3.2.2 The Common Ground in our brain

The processes involved in text comprehension described in (1) were investigated in a combined ERP and fMRI study by Schmalhofer et al. (2005). The results allowed them to distinguish separate brain processes such as memory resonance (see (1c) above) and situational constructions (like the creation of situation models from mental entities, propositions and inferences, (1d) above). Later behavioural studies by, among others, C. L. Yang, Perfetti, and Schmalhofer (2007), point to the same results, separating the ERP-effects in even more detail. There is thus psycholinguistic evidence for the cognitive situation model as described above.

The *communication* model of the Common Ground (CG) discussed before contains both entities and mutually accepted propositions (cf. Krifka (2008)). The Common Ground is constantly updated: new entities and propositions are introduced as the discourse moves along. The propositions are not only derived from the discourse, but can also stem from common belief and world knowledge the interlocutors or readers have stored in memory. What Krifka (2008) describes as the Common Ground thus closely resembles the descriptions of the mental entities and propositions we use to build the situation or mental model in our brain, as we saw in the previous section. If this is indeed the case, the communicational model of the Common Ground has a cognitive correlate and at least some processes involved in information packaging, such as anaphora resolution (Van Berkum et al., 2007), foregrounding of information (Zwaan & Radvansky, 1998), topic identification (Kintsch & Rawson, 2005) or focus structures (Cowles, Walenski, & Kluender, 2007) can be measured by non-invasive studies of the brain.

The present study aims to describe Welsh information-structural processes and how they interact with the observed word order variation. As such, psycho- and neurolinguistic experiments that could further investigate the suggested correlation are beyond the scope of the present research. More detailed studies of IS and the Common Ground in many different languages can, however, certainly provide both inspiration and specific guidance concerning experimental settings that could show precisely how the communicational model of the Common Ground functions in our

brain.

3.2.3 CG content vs. CG management

So far we have mainly focussed on the *content* of the Common Ground, the set of entities and propositions that are known to and shared by the interlocutors or readers. Apart from this notion of CG content, Krifka (2008) introduces 'CG Management' for the way the CG content should develop. The CG management too is shared, but the responsibility for it "may be asymmetrically distributed among participants" (Krifka, 2008:17). This distinction between CG content and CG management can be observed in two different kinds of focus constructions that are called semantic or pragmatic focus respectively (cf. Krifka (2008:21)). Semantic focus is concerned with the factual information of the CG content; it can thus affect the truth-conditional content. Pragmatic forms of focus constructions serve the communicative goals of the participants and do not immediately influence the truth conditions. In section 3.3.4, I will get back to this division of the Common Ground with further explanation and examples of both types.

3.3 Coding Information Structure

In the previous chapter I discussed the technical side of developing an annotated database of historical Welsh. The texts are first of all digitised, PoS-tagged and chunkparsed and converted to xml-files to facilitate any queries into morphological or syntactic aspects. In addition to that, any information that could be relevant to information structure is added to each clause in the form of features rendering attribute-value pairs that are searchable as well (cf. Chapter 2). The following sections are concerned with these coded IS features. Which features were coded? Why those features and not others? And, finally, how were they coded? Which possible values belong to the feature attributes and how did I decide for one value or the other?

This chapter does not aim to provide an exhaustive overview of all IS terms and how they are used in the literature. Instead, it describes the strategies and definitions used in the present historical investigation of Welsh information structure. As a starting point, I assume that the information structure of every clause can be described as one of the following 'focus domains' or 'focus articulations' (cf. Lambrecht (1994) and Komen (2013), among others):

(5) a. THETIC focus (containing thetic and presentational sentences)
b. PREDICATE focus ('wide focus', 'information focus' or 'topic-comment' structure)
c. CONSTITUENT focus ('narrow focus' or 'identificational focus')

Lambrecht (1994) built on work by Gundel (1974) and Prince (1981) arguing that languages can focus three domains: the whole clause, the predicate of the clause or just a single constituent. In thetic sentences, both the subject and the predicate are

in focus (cf. Bailey (2009) and section 3.3.2). Predicate focus is the most frequently found focus domain, especially in narratives. It provides (new) information on an already established topic and is therefore often called 'topic-comment' structure (see section 3.3.3). Finally, in constituent focus one constituent is selected to be put against the background that forms the rest of the clause. The numerous ways of doing this will be discussed in section 3.3.4 below.

Both the referential state of the core arguments (see section 3.3.1) as well as syntactic and text-organisational (see Chapter 2) features help define the focus domain of the clause (cf. Komen (2013)). Two further pragmatic phenomena interact with each of the above-mentioned focus domains: the point of departure (or 'delimitation' or 'frame setting') and the principle of natural information flow (see section 3.3.6). Since the core arguments of copular clauses have a different syntactic configuration, I will discuss their information structural status separately in section 3.3.5.

The suggested IS annotation scheme thus covers multiple levels ranging from the referential state of the core arguments to the focus articulation of a clause, frame setting on a sentence level and discourse development in terms of cohesion of multiple sentences and paragraph/episode boundaries.

3.3.1 Given vs New: Referential State

> *"The origin of bees is from paradise and because of the sin of man they came thence; and God conferred his grace on them, and therefore the mass cannot be sung without the wax."*
>
> (Translation of *The Laws of Hywel Dda* by Wade-Evans (1909))

As we have seen in section 3.2 above, when we read a story we continuously add new entities and propositions to the Common Ground (cf. Chapter 2 of Komen (2013)). The to-be-added entities are first matched with whatever is part of the Common Ground already. If there is a perfect match with an existing entity in the CG, the features of the new phrase will be added to the existing mental entity, which is considered to be exactly identical. In the above fragment of a Welsh law text, for example, *bees* are introduced as a new entity and added as such to the CG. The third-person plural pronoun *they* a bit further on refers to the exact same entity as the *bees* that are just mentioned so they form a perfect match. The proposition in which the pronoun *they* occurs, the fact that *they came thence*, is now added to the mental entity we already created in the CG for *bees*.

But what about *paradise*, *the sin of man* and *God*? Neither of those were mentioned in the previous context, but we know nonetheless what they refer to. These entities are not identical to anything we previously added to the Common Ground. There is no textual antecedent; in other words, the denotations are assumed to be part of the 'world knowledge' of those living in a Christian society at least. Therefore they are stored in our long-term memory. This is exactly why the definite article can be used in the phrase *the sin of man*. We are not talking about a random sin. This is *the* sin everyone knows about: the reason man and, according to this Welsh

law, also bees, had to leave paradise. The definite article in *the mass* is there for the same reason: this concept is assumed to be known by the reader and is therefore not a completely new piece of information. The final phrase *the wax*, however, is not necessarily part of the assumed Christian model in our minds. Furthermore, when we try to match this with the existing entities in the Common Ground, we fail to find an exact match. The first entity we added (*bees*), however, evoked a link to a model of bees that we store in our long-term memory (e.g. bees are insects, they fly and buzz, they make honey, etc.). We can easily infer the existence of *wax* from the *bees* we already have in our Common Ground, so *the wax* in this example does not convey completely new information either.

This brief interlude about the importance of bees in Welsh laws serves as an introduction to one of the most crucial dimensions of information structure: givenness. From the early days of research into information structure, 'givenness' in its various forms has played a crucial role. The degree of Communicative Dynamism, as Firbas (1964) called it, is what pushes communication forward. Chafe's (1976) cognitive theory distinguishing degrees of givenness was extended by Yule (1981), among others. And in more recent literature, 'givenness' is (the extent to which a particular phrase is) 'existentially entailed by the context' (cf. Zimmermann and Féry (2010:2) following Schwarzschild (1999)). Krifka (2008:37) defines it in relation to its presence in the Common Ground, and/or the degree to which the particular referent is present. The same gradient notion we already encountered identifying some constituents as 'not completely new' in the introductory Welsh law text is found in the definition by Traugott and Pintzuk (2008:64): "the degree to which a referent is represented as identifiable by the addressee/reader and is "hearer/addressee-old". Gregory and Michaelis (2001) distinguish givenness from what they call 'anaphoricity', which is concerned with textual reference only, rather than the hearer's cognitive status.

In theory, the givenness or information/referential state of any kind of discourse referent can be assessed, but for the purpose of the present thesis only the core arguments of the sentence will be annotated. The 'information status', as Götze et al. (2007) call it, reflects the retrievability of the referent: how difficult is it to find an antecedent? Is there an identical match, can we infer or assume its existence? or is the noun phrase we are currently adding to the Common Ground not linked to anything at all? As we have seen in the introductory analysis of the bee fragment, there must be more than a simple binary option of given vs. new.

To capture this gradience a wide variety of taxonomies and hierarchies were developed over the years: Prince's (1981) taxonomy of given-new information or information states of noun phrases elaborated and refined by Birner (2006) into discourse and hearer old-new distinctions, Riester, Lorenz, and Seemann (2010)'s detailed set combined with semantic information, Ariel (1999)'s accessibility marking scale, Gundel, Hedberg, and Zacharski (1993)'s givenness hierarchy or the tag sets for PROIEL (Haug, 2009) or Cesac's Pentaset (Komen & Los, 2012:21,23) (see Komen (2013:133-154) for a detailed overview and evaluation of each of those).

Komen (2013) shows that a combination of syntactic annotation and a small set of five referential state primitive suffices to capture all relevant degrees of givenness. In this thesis, I employ this same 'Pentaset' to enrich the core arguments in the Welsh historical database. Komen's primitives are very similar to the PROIEL tag set (Haug, 2009), Birner's discourse/hearer distinctions (Birner, 2006) and to those suggested by Götze et al. (2007) in their Linguistic Information Structure Annotation (LISA) guidelines, although the latter is unable to capture certain subtle differences concerning anchoring (see below).

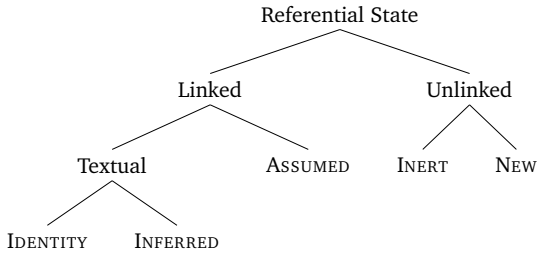
Taylor and Pintzuk (2014) test the effect of various annotation systems on Old English pre- and post-verbal objects. They find three significant differences: (i) between elaborating and bridging inferables, (ii) between specific new referents and short-term discourse referents and (iii) between short-term referents and semantically incorporated objects (Taylor & Pintzuk, 2014:72). As for (i), only Birner (2006) makes this distinction directly. In the Pentaset, however, the most-frequent cases of elaborating inferentials (the ones with inalienable possession) are marked with an Identity anchor (see discussion in the next section) and can thus be distinguished from bridging inferables. The next significant difference found between specific new referent, short-term referents and incorporated objects (numbers (ii) and (iii) above) fall in the Inert category in the Pentaset. They can be distinguished from other inert categories on the basis of their syntax and further featural annotation only. For the present study I used the Pentaset labels, because it makes more precise and clearer distinctions than the PROIEL or LISA annotations guidelines. In future research, it would be interesting to test Birner's (2006) distinctions on the Welsh dataset as well to see if there are similar significant results as the ones found for Old English object position by Taylor and Pintzuk (2014).[4] The main strength of the Komen's system is its ability to *derive* topic and focus structures from the IS and syntactic annotation combined. No additional assumptions have to be made to detect the right focus domain of a clause and it can even be extended to investigate copular clauses (the IS analysis of which is by my knowledge not specifically discussed elsewhere). In sections 3.3.3 and 3.3.4 below, I further develop the IS annotation system so that it can cover even more specific IS concepts such as the many different types of focus Krifka (2008) discusses, but also contrastive topics.

The Pentaset of referential state primitives

The referential state primitives that make up the pentaset are the minimal labels necessary to derive any other taxonomies or topic or focus domains (see Chapter 5 of Komen (2013) for a detailed overview). In this section, I provide definitions and examples for each of those five primitives. I furthermore point out subtle differences with the LISA guidelines by Götze et al. (2007). This is the Pentaset hierarchy (after Figure 11 in Komen (2013:144)):

[4] Taylor & Pintzuk's test results were published when the annotation with the Pentaset of the Middle Welsh database was already done.

(6)



The Pentaset is couched in the situation model (or Common Ground) discussed in section 3.2.1 above. The system first of all distinguishes noun phrases *with* an antecedent ('Linked') from those *without* ('Unlinked'). If there is a phrase (NP_i) referring to a certain mental entity $\texttt{MEnt}(NP_i)$ and there is another phrase (NP_j) that refers to the exact same mental entity of NP_i and NP_j linearly precedes NP_i, there is a perfect match with an already existing mental entity in our situation model. In this case, NP_i will receive an IDENTITY label, because its mental entity is identical to the mental entity of NP_j that already existed in our model. An example of this is a pronoun referring back to the mental entity created by a previously-mentioned NP. The formal definition of the IDENTITY label is, according to Komen (2013:144):

(7) **Identity**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ has the referential status "Identity" if there is an NP_j with $j<i$, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$.

The *bees* in the introduction that were matched by the pronouns *they* and *them* further on are a clear example of this. Götze et al. (2007) further divide this category, which they call 'given' into 'active' and 'inactive' referents. 'Active' referents are those that are referred to "within the last or in the current sentence" (Götze et al., 2007:154). There indeed seems to be a difference in terms of accessibility the further you move from the antecedent. The sentence boundary, however, is a somewhat arbitrary notion. In many medieval manuscripts, for example, it may be hard to divide the text into clear sentences in the first place. Clause boundaries are easier to define, but there can be multiple subordinate clauses in one sentence, so cutting off at one, two or three clauses or even one matrix clause remains a random decision. It remains unclear, however, whether "one sentence" is meaningful as an IS notion here.

Looking at the last-mentioned possible antecedent could be a more meaningful distinction, but even that may vary from language to language. Grammars can act differently if they have no (rigid) gender or number marking in the nominal system, for example, from those with 'rich' morphological paradigms of pronouns and demonstratives. I leave this as an open question for now, because for the present investigation, this particular distinction is not relevant. In the present thesis I will stick to the simple IDENTITY label for any referent that has an exact match with a mental entity that is referred to in the previous context. In long narratives

featuring the same main characters over and over again, I will furthermore indicate whether the specific referent occurred in the same *scene* or not. A change in location or setting is a clear indication of a scene boundary. If the hero of the story disappears for a while, for example, because the narrative changes its focus for a few paragraphs, we replace the model we created in our mind. The same hero can then be identified later on, but the scene has changed so the particular noun phrase will receive an additional label: IDENTITY - CHANGE OF SCENE as the subject *y mab* 'the boy' in this example following a scene in which the father of the boy gives his son advice on how to find Olwen:

(8) *Mynet a oruc y mab ar orwyd penlluchlwyt...*
go.INF PRT do.PAST.3S the boy on steed gleaming-grey-head...
'The boy went off on a steed with a gleaming grey head...' (CO 60)

Antecedents can occur in the text, but they can also be part of the general 'world knowledge' stored in our long-term memory. Entities in our long-term memory can be evoked and become part of the Common Ground. When this type of link to an entity in long-term memory can be created, the referential state of the mental entity that is added to the situation model is ASSUMED. Komen (2013:147) gives the following formal definition of the categorie ASSUMED:

(9) **Assumed**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ is "Assumed" if
a. there is no NP_j with j<i, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$
b. nor such that $\texttt{MEnt}(NP_j)$ can be inferred from $\texttt{MEnt}(NP_i)$, but
c. there exists an $\texttt{MEnt}(NP_{LTM})$ (in long-term memory), such that $\texttt{MEnt}(NP_{LTM}) = \texttt{MEnt}(NP_i)$

We have seen examples of this in the fragment on the origin of the bees above: *God*, *paradise*, *the mass* and even *the sin of man* do not need a textual antecedent to be meaningful to a reader who is familiar with at least the basic background of the Christian faith. This is considered 'world knowledge', just as much as we all know the sun, moon and stars exist. Situational knowledge about the speaker, hearer, the book that is being written or the setting in which the sentence is uttered, also belongs in this category. Imagine, for example, a conversation over lunch where one person points to the box on the other side of the table and asks:

(10) "Could you pass the chocolate sprinkles, please?"

The noun phrase *the chocolate sprinkles* has an antecedent, even though it was not mentioned in previous discourse. The other person can see the box of chocolate sprinkles on the table, so the new mental entity of the noun phrase will match the referent in the currently relevant situation. The referential state of the phrase *the chocolate sprinkles* is thus ASSUMED. In the extended tag set of the LISA guidelines, Götze et al. (2007) create a special label for referents that are part of the discourse situation such as *the chocolate sprinkles*: 'accessible-situative'. Since it is unclear if

and why matching with something in our long-term memory or with the situation at hand would make a difference, I will stick to the Pentaset label for referents whose information status is ASSUMED.

It can also be the case that there is no direct match with a textual antecedent or an antecedent in the current situation or long-term memory, but the information referred to is not completely new either. In the introductory fragment, *the wax* was an example of this, because we could establish a link with the afore-mentioned *bees* via the model concerning bees in our long-term memory that was evoked as soon as we read about them. In other words, we could infer the existence of *the wax* from the *bees*. When this form of logical reasoning is necessary to establish an entity, Komen (2013:146) defines it as INFERRED:

(11) **Inferred**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ has the referential status "Inferred" if
(i) there is no NP_j with $j<i$, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$, but
(ii) there is an NP_k with $k<i$, such that:
a. $\texttt{MEnt}(NP_i) \in S_x$
b. $\texttt{MEnt}(NP_k) \in S_y$
c. there exists a *direct set relation* between set S_x and S_y.

A direct set relation, as used in this definition can for example occur in the form of a subset, a part-whole relation or as an entity-attribute relation as in the following examples:

(12) a. Deryn hates working close to the microwave. *The noise* is distracting.
b. Asiye loved the Turkish chocolates. *Their flavour* was so soothing.

The italicised noun phrases in examples (12a) and (12b) create mental entities that are not *identical* to anything in our situation model or in long-term memory. We can, however, create a link to the existing entities, *the microwave* and *the Turkish chocolates*, because there exists a direct set relation: microwaves make a lot of noise and chocolates have flavours. If there is no antecedent in the context (IDENTITY), in our long-term memory or direct situation (ASSUMED) and if we cannot infer the existence of the referent from anything previously mentioned (INFERRED), the referential state of the phrase is 'Unlinked'. The Pentaset further differentiates the 'Unlinked' category: referential phrases that could serve as an antecedent in the following discourse are labelled NEW (Komen, 2013:150):

(13) **New**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ is "New" if
a. there is no $\texttt{MEnt}(NP_j)$ with $j<i$, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$,
b. nor such that $\texttt{MEnt}(NP_j)$ can be inferred from $\texttt{MEnt}(NP_i)$, but
c. it **is** possible that there exists an NP_k with $k>i$,
such that $\texttt{MEnt}(NP_k) = \texttt{MEnt}(NP_i)$.

New entities are usually introduced as indefinite noun phrases or phrases with postmodifiers, as in the following examples:

(14) a. *Ac yno ti a wely lwyn.*
and there you PRT see.2S grove
'And there you will see a grove.' (Peredur 294)
b. *A ffon yssyd idaw o hayarn*
And stick be.3S to.3MS of iron
'And he has an iron stick.' (WM 228.23-24)

There are also phrases that can *not* be referred to in the following context. Usually, they function as attributes of other entities. Götze et al. (2007) do not have a specific label for these expressions in the LISA guidelines, exactly because of this reason: they do not annotate "NPs or PPs that don't refer to discourse referents". Examples of non-referential expressions are expletives or parts of idiomatic phrases or attributes as in:

(15) a. Mabon son of Modron is here in *prison*; and none was ever so cruelly imprisoned in a prison house as I.
b. Maxen Wledig was emperor of Rome, and he was *a comelier man*.

In example (15a), the prisoner Mabon is shouting from within his confined space in a cry for help. The noun phrase *prison* in the first part of the sentence refers to the general concept of his confinement. The phrase is INERT: it cannot serve as an antecedent for the following discourse. Similarly, in example (15b), the noun phrase *a comelier man* cannot be picked up later on. A following sentence starting with *The man went hunting.* sounds odd at the very least (cf. Johnson-Laird (1983) and Komen (2013)).

To sum up this section, I give a full analysis of the referential status of the most important noun phrases in the following fragment from the translation of *Culhwch ac Olwen*, the oldest Arthurian tale. The immediately preceding context relates how Arthur was hunting a wild boar called Twrch Trwyth. The boar has fled to Ireland and Menw tried to capture it, but failed, upon which Twrch Trwyth destroyed a large part of the country. There is a brief intermezzo about a magic cauldron and then...

(16) ***Arthur*** *came to* ***Esgeir Oerfel*** *in* ***Ireland,***
to the place where ***Twrch Trwyth*** *was,*
and ***his seven young pigs*** *with* ***him****.*

Arthur is one of the main characters of the tale and is also mentioned in the immediately preceding context. The referential state is thus IDENTITY. *Esgeir Oerfel* on the other hand, is NEW in this context. It was mentioned once or twice in the beginning of the tale, but since there were many different scenes in between and this place does not play any significant role in the tale, it is unlikely that this is still in our situation model. If this was a famous place in Ireland, a medieval Welsh audience might have stored it in their long-term memory, rendering its referential

state ASSUMED. As for *Ireland* itself, this too was mentioned in the immediately preceding context, so this, just like *Twrch Trwyth* and *him* at the end of the first sentence, is labelled IDENTITY. Finally, *his seven young pigs* bring a new entity into our situation model, because these pigs were not mentioned before. The phrase is linked to the wild boar Twrch Trwyth in two ways. First of all, we can establish an inferential relation between pigs and wild boars, because boars (can) have pigs. The referential state will thus be INFERRED. But there is another element in the phrase linking it to this wild boar in particular: the possessive pronoun *his*. Following Prince (1981) and Komen (2013), I call this an "identity anchor". This anchor can be added independently: the full referential state of *his seven young pigs* will thus be INFERRED + IDENTITY ANCHOR.

(17) a. ***Dogs*** *were let loose at* ***him*** *from* ***all sides****.*
b. ***That day*** *until evening* ***the Irish*** *fought with him; (...)*
c. ***His men*** *asked* ***Arthur*** *what was* ***the history of that swine****, and he told them:*
d. *'He was* ***a king****, and for* ***his wickedness God*** *transformed him into a swine.'*

When we continue reading, we find *dogs* in (17a), which forms a NEW mental entity in our situation model, as opposed to the pronoun *him*, which forms a perfect match with the wild boar we have seen before and is thus labelled as IDENTITY. The phrase *all sides* is not linked to anything either, but this phrase does not add an entity to our Common Ground, because it is non-referential. It cannot serve as an antecedent in the following discourse, so we will label it as INERT. The first phrase in (17b), *that day* is ASSUMED, because it is part of the current situation. We can infer the existence of *the Irish* from the previously-mentioned *Ireland*: countries have inhabitants, a country called 'Ireland' has inhabitants that are called 'the Irish'. Its label is INFERRED. In (17c), *his men* form a new mental entity, because these men just come to the scene. There is a possessive pronoun, however, that links this phrase to *Arthur*. Therefore it will get the label NEW + IDENTITY ANCHOR, making it more accessible than new entities without any form of anchoring in the previous context. The same goes for *the history of that swine*: the *history* is NEW, but the *swine* is already well-established in our model, so this too gets the label NEW + IDENTITY ANCHOR. This is also the case for the phrase *his wickedness* in (17d). The phrase *a king* is not linked to anything, but again, it is very difficult to see how this phrase could serve as an antecedent. It is a clear example of an attributive indefinite noun phrase in the complement position of an equative clause and therefore INERT. *God*, finally, is an entity that we can link to a concept in our long-term memory and it is therefore labelled as ASSUMED.

3.3.2 Presentational or Thetic structures

Once we have annotated the morphology (see PoS-tagging in Chapter 2), basic syntactic structure (see Chunkparsing in Chapter 2) and the referential state (see section 3.3.1), we can derive the focus domain from this combined information

(Komen, 2013). Presentational or thetic sentences focus both the subject and the predicate. If the sentence merely consists of a comment and there is no core argument constituent that could be the topic, the sentence is called 'thetic' (cf. Krifka (2008:43) following Marty (1884)). I therefore label the clause's focus domain as THETIC FOCUS. Krifka (2008:43) gives the following example of a sentence without a topic constituent in a situation where somebody is running towards you in a panic, for example, shouting:

(18) [*The HOUSE is on fire.*]$_{\text{Comment}}$

There still is a topic *denotation* in this clause, the sentence is still 'about' something. But there is no constituent expressing this, because the entire sentence consists of a comment explaining someone's panic. Both the subject and the predicate, convey new information. Another example of a thetic statement is:

(19) [It is raining]$_{\text{Comment}}$.

The topic of sentences like (19) is also called a "stage topic" (cf. Gundel (1974) and Sasse (1987)), because it predicates about the 'here and now'.

Presentational sentences are similar in that they also contain subjects and predicates that are NEW. They are relatively easy to recognise, because they introduce a new entity into the discourse. Very often, these sentences occur at the beginning of narratives:

(20) *In the days when Maelgwn Gwynedd was holding court in Castell Deganwy, there was* ***a holy man named Cybi*** *living in Môn.*

In this opening passage of *Ystoria Taliesin* from the 16th-century Chronicle of the World by Elis Gruffudd, a new entity is introduced, namely *a holy man named Cybi*. The preceding prepositional phrase *In the days...* functions as a point of departure or 'frame setting' (see section 3.3.6), but the focus domain is determined by the rest of the clause in which a new entity is introduced as the subject. The focus domain of this clause comprises the subject and the predicate and it is thus labelled THETIC FOCUS as well.

Other examples of thetic focus will be discussed in section 3.3.5 below. For now it suffices to say the thetic focus domain can be detected when the subject contains NEW information and the predicate is also part of the focus domain. Komen (2013:42) furthermore adds that thetic focus can be overridden by constituent focus. This means that if the subject is, for example, providing the value for a variable that has just been raised, the sentence does not belong to the thetic focus domain, but receives the label of CONSTITUENT FOCUS (see also section 3.3.4). The example Komen (2013:42) gives is the following dialogue:

(21) a. "Who would want to listen to you?"
b. "*An educated man* will read my books!"

The italicised noun phrase in (21b) provides the value for the variable created by the question in (21a). The predicate *read my books* in (21b) furthermore does not contain completely new information, because the verb *read* can be inferred from *listen* in (21a) (cf. Komen (2013:42). In this case, the clause in (21b) is thus an example of a CONSTITUENT FOCUS domain. Apart from CONSTITUENT and THETIC FOCUS, there is a third type of focus domain called PREDICATE FOCUS for topic-comment structure. This domain is discussed in the next section.

3.3.3 Topic vs. Comment

Consider the following fragment from the tale of Branwen, the second branch of the *Mabinogion*, translated by Lady Charlotte Guest and try to think of what this passage is *about*:

> *"In Ireland none were left alive, except five pregnant women in a cave in the Irish wilderness; and to these five women in the same night were born five sons, whom they nursed until they became grown-up youths. And they thought about wives, and they at the same time desired to possess them, and each took a wife of the mothers of their companions, and they governed the country and peopled it. And these five divided it amongst them, and because of this partition are the five divisions of Ireland still so termed. And they examined the land where the battles had taken place, and they found gold and silver until they became wealthy."*
>
> (Guest, 1849)

The most logical answer is that it is about five sons who grew up to 'people' Ireland: that is the topic of this piece of discourse. There is a vast literature on different kinds of topics including various definitions, functions and ways to express them. In this section, I discuss only those notions relevant for the present thesis. Starting with a definition of topic by Krifka (2008) (following Reinhart (1981)), I continue to characterise the most frequently found focus domain called PREDICATE FOCUS that consists of the basic topic-comment structure and whose frequent occurrence in narratives makes sense from a cognitive point of view. Finally, I describe different kinds of topics in sentences and discourse and how they can be marked in the grammar.

Topics and the Predicate focus domain

Krifka (2008:41) defines topic constituents in the following way:

(22) "The topic constituent identifies the entity or set of entities under which the information expressed in the comment constituent should be stored in the CG content."

The content of the Common Ground thus plays a crucial role. The propositions in the CG are stored under certain entities just like the file card system proposed by Reinhart (1981) and Vallduví (1992). Other definitions of 'topic' containing 'subject' (cf. Chafe (1976)) or 'theme' conflated with 'old information' (cf. the Prague

School, e.g. Daneš (1970)) should according to Krifka (2008) and Zimmermann and Féry (2010) be avoided, because they are not necessarily grammatical subjects or inferable from the preceding context.

There are various ways to find topics described in the literature. Gundel (1988:210)'s definition comprising the speaker's intention "to increase the addressee's knowledge about, request information about, or otherwise get the addressee to act with respect to" the topic of the sentence is an intuitive working definition, but it does not give any concrete guidance on how to identify topics. Götze et al. (2007:165) formulate three conditions identifying aboutness topics X in sentence S if:

(23) a. S would be a natural continuation to the announcement: "Let me tell you something about X."
b. S would be a good answer to the question: "What about X?"
c. S could be naturally transformed into the sentence "Concerning X, S'." or into the sentence "Concerning X,S'," where S' differs from S only insofar as X has been replaced by a suitable pronoun.

Eckhoff and Haug (2011) are more precise and formulate an algorithm that ranks constituents that are possible topic candidates according to parameters such as their referential status, animacy, morphosyntactic realisation, saliency, syntactic relation, word order and antecedent properties. The strength of this algorithm lies in the combination of those features yielding 90% agreement between the outcomes of their algorithm and that of human intuition. In a similar way, the Cesac application (Komen, 2009a) attempts to detect topics based on the type of NP and their grammatical function (subject, object, etc.). Centering theory finally, (cf. Grosz, Weinstein, and Joshi (1995), and in particular the OT type of centering discussed by Beaver (2004)), is according to Komen (2013) a particularly successful way to find the topic of a sentence. It ranks the topic candidates according to their category (e.g. demonstrative, pronoun, definite noun phrases, etc.), the referential state of the phrase (linked or not) and their grammatical role (e.g. subject or object).

In the present research, all these notions (and more) are annotated in the database to facilitate the search for topics in each sentence, separating them from the rest of the clause that makes up the comment. In terms of focus domains, this topic-comment structure differs from the above-mentioned THETIC sentences in the sense that the latter always contain subjects (and predicates) conveying new information: both subject and predicate are in focus. In topic-comment structures, the focus domain is the predicate that conveys the NEW information. This is also called 'wide' or 'information focus' (e.g. É.Kiss (1998)), but following Lambrecht (1994) and Komen (2013), I label this domain PREDICATE FOCUS.

Why exactly is this type of focus domain the one we find most frequently in narratives? Psycholinguistic experiments (e.g. Gernsbacher (1990)) have shown that from a processing perspective, the predicate focus domain with the topic-before-comment structure is likely to be the most commonly used, since language

is processed in a largely incrementally way. It furthermore makes sense to present linked information before unlinked information in the predicate. In this respect, it is also interesting to look at VOS and, in particular, VSO languages and their topic-comment distribution, because the verb in the latter case is fronted leaving the direct object behind and thus the focussed predicate is split up (more on Modern Welsh VSO is discussed in Chapter 5). As Cowles (2012) puts is: "when we encounter or produce a sentence we begin to process it right away, at the beginning, without waiting for the entire sentence to be available for either production or comprehension." (Cowles, 2012:290). This first information to be processed is very often the given referent, but there is also evidence from German that sometimes new information may be ordered first (cf. Cowles (2012)). For the present research, it suffices to say that two IS notions that seem particularly relevant in topic-comment structures, namely givenness (see section 3.3.1) and accessibility (see Chapter 2) are annotated separately. If topic-status is, as these production studies indicate, indeed assigned at the pre-linguistic message level, we need to investigate how this can be encoded in the grammar in general. In this thesis I show how this can be done in earlier stages of the Welsh language and how this changed over time.

Finding the focus domain

PREDICATE FOCUS is the most frequently found focus domain in narratives, as we have seen in the introductory fragment about the five sons. Every predicate of the following sentence adds new information, a new file-card if you will, to the existing entity: the sons want wives, get married to each other's mothers, govern the country, etc. We can find this focus domain of the sentence by following a decision-making tree based on the combined syntactic and referential state information of the core constituents of the matrix clause. It is also possible to determine the focus domain of subordinate clauses (see Chapter 2), but here we try to determine the focus domain of matrix clauses first.

First of all, we make sure we are not dealing with a thetic or presentational sentence by asking the following questions:

(24) Is there a topic constituent?
  (i) Yes ⇝ Move on to (25)
  (ii) No ⇝ Are both subject and predicate new?
    (i) Yes ⇝ THETIC FOCUS
    (ii) No ⇝ Start over (something went wrong).

(25) Is there a new entity introduced into the story?
  (i) Yes ⇝ THETIC/PRESENTATIONAL FOCUS
  (ii) No ⇝ Move on to (26)

After ruling out the domain of THETIC FOCUS, we check if we are dealing with a copular clause (see section 3.3.5). If this is not the case we continue to ask whether the sentence forms part of a dialogue with a whole set of further questions to rule out various types of CONSTITUENT FOCUS (see section 3.3.4 below). If the sentence

is not part of a dialogue, we first of all see if this is a case of a contrastive topic (see section on Types of Topic below). Finally, we distinguish between the domains PREDICATE and CONSTITUENT focus by asking whether there are relevant alternatives for any of the constituents in the clause, based on Krifka (2008)'s definition of focus (see section 3.3.4 below). If this is not the case, we are almost certainly dealing with a topic-comment structure and label it PREDICATE FOCUS. We can furthermore test this by finding the topic (combining different pieces of information as described above) and establishing the referential state of the predicate. If the predicate adds new information to the topic in a file-card manner, we are indeed dealing with the most commonly found focus domain: PREDICATE FOCUS. Schematically, this procedure looks as follows:

(26) Is it a copular clause?
  (i) Yes ⇝ Go to copular clauses (see section 3.3.5)
  (ii) No ⇝ Is it part of a dialogue?
    (i) Yes ⇝ Go to dialogue options (see section 3.3.4)
    (ii) No ⇝ Is there a contrastive topic?
      (i) Yes ⇝ PREDICATE FOCUS + CONTRASTIVE TOPIC
      (ii) No ⇝ Are there relevant alternatives for one of the constituents?
        (i) Yes ⇝ CONSTITUENT FOCUS (see section 3.3.4)
        (ii) No ⇝ PREDICATE FOCUS

The type of CONSTITUENT FOCUS will be specified in section 3.3.4 below. But with the above decision making tree, we can determine the domain of focus in every clause: THETIC, PREDICATE or CONSTITUENT FOCUS.

Types of topics

Topics come in different kinds and shapes. In the previous section, we zoomed in on the most common type, the 'aboutness topic'. This is also the kind of topic that is usually meant in IS literature (although it differs from the 'syntactic topic' in studies of the information structure of Old English, which denotes the first constituent of the sentence, cf. Traugott and Pintzuk (2008:64)). Götze et al. (2007) furthermore have a special label in the LISA guidelines for what they call 'frame-setting' topics that "constitute the frame within which the main predication of the respective sentence has to be interpreted." (Götze et al., 2007:167) and they give the following example:

(27) *Körperlich geht es Peter sehr gut.*
Physically goes it Peter very well.
'Physically, Peter is doing very well.' (German)

The frame setter in this sentence is the adverb *körperlich* 'physically', but the sentence also has an aboutness topic, namely *Peter*. Götze et al. (2007) choose to annotate both topics in this case, one as an 'aboutness' topic and the other as a 'frame-setting topic'. I chose to treat these frame setters differently labelling

the sentence as having POINT OF DEPARTURE (cf. Komen (2013:44-46) and section 3.3.6 below), because these frame setters interact with all three types of focus domains and do not exactly function like the 'aboutness' topics. According to Krifka (2008:46), for example, frame setters can indicate "the general type of information that can be given about an individual". He interprets frame setters as delimitators restricting the notions that can be expressed to the indicated dimension of a clause, e.g. as for his physique / physically, in example (27). The crucial point of frame setters is the possibility of alternatives, which makes them always focussed in a sense, following from Krifka (2008)'s definition of focus (see section 3.3.4 below). There would be no need for a frame setter in the first place, if there is no alternative perspective: they imply that "there are other aspects for which other predications might hold" (Krifka, 2008:46). As such, they behave similarly to what Büring (2003) and Krifka (2008) have called "contrastive topics". Contrastive topics are "topics with a rising accent" representing "a combination of topic and focus" (Krifka, 2008:44). Just like frame setters, they can take a complex issue and split it into sub-issues. Consider first Krifka's (2008) example from an English dialogue in (28):

(28) A: What do your siblings do?
B: [My [SISter]$_{\text{Focus}}$]$_{\text{Topic}}$ [studies MEDicine]$_{\text{Focus}}$,
and [my [BROther]$_{\text{Focus}}$]$_{\text{Topic}}$ is [working on a FREIGHT ship]$_{\text{Focus}}$.

The two topics are contrastive in (28), but they really function as the topic with new information added in the focussed predicate. The rising accent indicated with the capital letters furthermore denotes some sort of focus to show the contrast as a strategy of incremental answering in the CG management. In Middle Welsh, we do not have the necessary information about accents, but we do find examples that look very similar. The first example is found in a passage in the Welsh Laws describing the rights of the officers of the court; the second is from the Middle Welsh Arthurian tale *Culhwch ac Olwen*:

(29) a. *[Brenhines]*$_{\text{Topic}}$ *a geif [trayan gan y brenhin]*$_{\text{Focus}}$ *(...), ac velly*
queen PRT get third by the king (...) and so
y dyly [sswydogion y vrenhines]$_{\text{Topic}}$ *[y trayan gann swydogion*
PRT entitled officers the queen the third by officers
y brenhin]$_{\text{Focus}}$.
the king
'The queen will get a third from the king (...), and so the officers of the queen are entitled to a third from the officers of the king.'
(Cyfreithiau Hywel Dda yn ôl Ll. BL Add. 22356, 5.11)

b. *[Y trywyr]*$_{\text{Topic}}$ *a [ganant eu kyrn]*$_{\text{Focus}}$*, a ['r rei ereill*
the three.men PRT play.3P their horns and the some others
oll]$_{\text{Topic}}$ *a [doant y diaspedein]*$_{\text{Focus}}$
all PRT come.3P the outcry
'The three men shall play their horns, and all the others will come to make outcry.'
(CO 743-744)

In such cases where there is a clear contrast between two aboutness topics in one sentence that has another focus (e.g. in the predicate in the above examples), I label them as CONTRASTIVE TOPIC.

This extra focus outside the topic, also holds for the frame setters. In an attempt to capture this delimitating function of both frame setters and contrastive topics, Krifka (2008:48) characterises these structures as follows:

(30) A Delimitator α in an expression [...α...β_{Focus}...] always comes with a focus *within* α that generates alternatives α'. It indicates that the current informational needs of the CG are not wholly satisfied by [...α...β_{Focus}...], but would be satisfied by additional expressions of the general form [...α'...β'$_{\text{Focus}}$...].

This definition allows for more types of delimitators than the two mentioned here, contrastive topics and frame setters. It might, however, be too strict to include examples like (29a) and (29b). Without access to prosodic information, it is hard to establish whether there would be a rising accent, for example, and thus focus on the topics *brenhines* 'queen' and *sswydogion y vrenhines* 'the officers of the queen'. In order to let them count as real examples of **Delimitation**, according to Krifka (2008), we would have to assume the CG is not 'wholly satisfied' without the second part of the sentence. It is not altogether clear whether this is the case, because 'The queen will get a third from the king' could make perfect sense in itself in a law text that describes the legal rights of the queen. If there is evidence to the contrary, e.g. because from the context it is clear that the sentence is not complete without the second clause, example (29a) would indeed count as a Delimitator under Krifka's definition.

In the context preceding example (29b), the giant Ysbadadden Pencawr lists a number of men and beasts that are required to hunt the wild boar, Twrch Trwyth (see also example (16) above). He then specifies what the three men will do: they will blow their horns. All the others he mentions will then come and cry out. Here too, we could argue that we expect the second part of the sentence: we do not just want to know what the three men of the long list will do, we also want information about the others.

Since it seems difficult to apply the general notion of Delimitation in historical data where we have no access to prosodic information, I have annotated examples like (29a) and (29b) and those with explicit frame setters on the basis of what we *can* detect from the sentence and the context. Frame setters will receive a POINT OF DEPARTURE label with a further specification according to their function (see section 3.3.6 below); topics that are contrasted with a topic in the following clause, with separate focus structures in the predicate as we have seen above, are labelled CONTRASTIVE TOPICS. I leave aside the question here whether contrastive topics are aboutness topics as well. Evidence from parallel (gapping) structures indicates that this is not necessarily the case (cf. Repp (2010)). This distinction is, however, not relevant for the present thesis.

In some historical studies (e.g. Frascarelli and Hinterhölzl (2007) and Walkden (2014)), a further distinction is made between 'Aboutness' and 'Familiar' Topics.

FAMILIAR TOPICS are D-linked topics (i.e. linked to an antecedent in the preceding discourse) that occupy a lower position in the left periphery of the clause than ABOUTNESS TOPICS. In Middle Welsh, only one argument can occupy a pre-verbal position. Only if the ABOUTNESS TOPIC acts as a frame setter (e.g. a temporal or locational phrase), can we find a second topic that could be labelled as the FAMILIAR TOPIC. In Chapter 7, I discuss this difference further in the context of the Middle Welsh Abnormal Sentences.

Topics in discourse

> *"The maiden came inside. 'Maiden,' he said, 'are you still a maiden?' 'I know no reason why I should not be.' Then he took the magic wand and bent it. 'Step over this,' he said, 'and if you are a maiden, I will know it.' Then she stepped over the magic wand, and in that step she dropped a large boy with curly yellow hair. What the boy did was give a loud cry. After the boy's cry, she made for the door, and in the process a little something dropped from her."*
>
> (Parker, 2007)

As we have seen in the fragment about the five sons in Ireland in the previous section, aboutness topics can be the center of attention for a longer period, extending beyond one single sentence to paragraphs, texts or complete conversations. This is not the case, however, in the above fragment from Math (the fourth branch of the *Mabinogion*), because first we focus on the maiden (and her virginity test; the Welsh text uses the same word for 'maiden' and 'virgin' here, hence this translation by Parker). After that we switch to the boy that dropped out of her, only to go back to the maiden again when she is making for the door.

In the field of discourse studies, much work has been done on identifying "topic chains" or "focus chains" (Erteschik-Shir, 2007:3). Topics can be derived or introduced in three ways: a) from the topic of the previous clause ("topic chain"), b) from the rheme of the previous clause ("focus chain") or c) from a hypertheme (cf. Daneš (1974)). Topic chains or 'topic persistence' is simply the continuation of the same topic in the following sentence(s). Traugott and Pintzuk (2008:70) distinguish this from "Subsequent Mention". Subsequent Mention requires that the topic constituent is referred to again, as opposed to "Topic Persistence" indicating a continuity of pragmatic/aboutness topics. In the above fragment, *the magic wand* is brought up and subsequently mentioned in the next sentences, but the *maiden* is the topic of the following sentence where she steps over the wand, not the magic wand itself. The topic chain is broken up by the boy that dropped out of her while she steps over the magic wand. From the rheme or focussed part of this sentence, the boy is taken as the topic of the next sentence where he gives a loud cry, thus forming a "focus chain".

According to Daneš (1974), a topic can also be derived from a "hypertheme". This hypertheme consists of a set of elements restricted by the discourse. Erteschik-Shir (2007:3) gives the following example:

(31) I'll tell you about my friends, *John, Paul, and Mary*. *John* is an old friend from school, *Paul* I met at college, and *Mary* is a colleague at work.

The topics in examples (31) above can be derived from a hypertheme that explicitly mentions all members of the set, as in (31), or it can describe the set, as long as its members are obvious. The distinction between topic and focus chains could be derived from the annotated historical Welsh data automatically. Hyperthemes are not marked as such, but the referential state INFERRED of a particular entity indicates a set relation nonetheless. The final part of this section concludes the discussion on PREDICATE FOCUS with an overview of how topics can be marked in the grammar of a (written) language.

Marking topics

Topics can be marked in various ways. Since the use of specific lexical items to mark topics is not relevant in the Welsh language, I will not discuss this option further here. Prosody and intonational patterns are notoriously difficult to investigate in historical sources. If the boundaries of prosodic phrases consistently coincide with syntactic phrases and if we know more about stress and metrics, we can start looking at prosodical patterns relevant for information-structural categories. This has been done, for example, for Old High German by Hinterhölzl (2009). Since our knowledge of this in Middle or Early Modern Welsh is still limited, for now I focus on those IS markings we *can* observe in our data, for example, the word order.

Word order and 'fronting' in particular has received much attention in the literature about information structure and topicalisation. 'Fronting' is a general term for the leftward movement of a constituent that is 'topicalised', i.e. put in a position where it is interpreted as the topic of the sentence. In West-Germanic languages like German, Dutch (dialects) or Frisian with a verb-second constraint in matrix clauses, topicalisation can be implemented in three ways: movement of a constituent (an NP or even an entire clause) (see (32)), left dislocation (see (33)) or as a hanging topic (see (34)):[5]

(32) **Movement**

a. *Diesen Mann habe ich noch nie gesehen.*
this.ACC man have I yet never seen
'I have never seen this man.' (German)

b. *De zon in oew leve kan ik oe nie geve.*
the sun in your life can I you not give
'I cannot give you the sun in your life.'
(Brabantish, from *Lieke vur Mariken* by Gerard van Maasakkers)

[5] According to Ross (1986:253n18), the term 'left dislocation' was coined by Maurice Gross. The term 'hanging topic' was, according to Cinque (1977:406) coined by Alexander Grosu.

(33) **Left Dislocation**

a. *Den Hans, den kenne ich seit langem.*
the.ACC Hans this.ACC know I since long
'Hans I've known for a long time.' (German, Cardinaletti, Cinque, and Giusti (1988:9))

b. *Di lieke, da zing ik vur jou.*
this song that sing I for you
'I sing this song for you.'
(Brabantish, from *Lieke vur Mariken* by Gerard van Maasakkers)

(34) **Hanging Topic**

a. *Der Hans - ich kenne diesen Kerl seit langem.*
the.NOM Hans - I know this.ACC guy since long
'Hans - I've known this guy for a long time.' (German, Nolda (2004:424))

b. *Skulpen, troch de ieuwen hinne hawwe minsken dy al sammele.*
shells through the centuries through have people them already collected
'Shells, throughout the centuries people have collected them.'
(Frisian, from `http://pers.tresoar.nl/bericht.php?id=377`)

The main difference between sentences like (32) labelled 'movement' and sentences with left dislocation of a constituent or a 'hanging topic' can be detected from the prosodic structure: in (33) and (34) the commas clearly indicate a pause separating the fronted constituent from the rest of the sentence. A further difference between (33) and (34) can be observed in languages with morphological case marking like German. Sentences with hanging topics are therefore also called 'nominativus pendens'.

According to Willis (1998), Middle Welsh also had a verb-second constraint. Consider the following example with a fronted direct object:

(35) *Ac ystryw a wnaeth y Gwydyl*
and trick PRT made the Irish
'And the Irish played a trick.' (Middle Welsh, PKM 44.11)

Why is the direct object constituent fronted in (35)? What is its exact referential status? What is the information structure of this clause and how does it fit in the context? One of the main research questions of the present thesis is concerned with the variation in word order and to what extent, if at all, this relates to information-structural features. To investigate this properly, we have to take all possible IS features into account. The syntactic and clause type features were discussed in Chapter 2, all other IS notions and their annotation are discussed in this chapter.

In Chapter 4 and 5, I zoom in on the historical Welsh data and the main generalisations concerning the interaction of IS and word order. One important question is, for example, if all above-mentioned 'fronting' or topicalisation strategies

are found in Middle and Early Modern Welsh, what their exact IS status is, and possibly how and why this changed over the centuries. Middle English had a verb-second rule with topicalisation strategies, but this is no longer found in present-day English (cf. Holmberg (2013)). Middle Welsh and closely related Middle and Modern Breton have a verb-second constraint, but the word order of Modern Welsh (VSO) is very different from present-day English (SVO). These issues and their interaction with topicalisation strategies are discussed in the following chapters.

3.3.4 Focus vs. Background

> *Gwen Cooper: 'That was your last chance!'*
> *Lyn Peterfield: 'Yeah? What are you going to do about it? If you're the best England has to offer, God help you!' [Silence while Gwen gets up.]*
> *Gwen Cooper: 'I'm WELSH.' [And Gwen punches her out.]*
> (scene from BBC's Torchwood, season 4, episode 2)

Focus is as much an intuitive notion as it is a linguistic one. Intuitively, or generally, we are inclined to associate 'focus' with 'contrast' as in the above dialogue, or 'emphasis' of some sort. This latter part is exactly what makes focus so difficult to define linguistically. A definition of focus comprising 'emphasis' requires a strict definition or a description of 'emphasis' at the very least. In an attempt to capture all different types of focus, linguistic notions vary from a general 'new' (versus 'given', 'background' or 'presupposed') information to more specific contrastive (versus non-contrastive) information. The notion of contrast is, however, not necessarily limited to focus constructions, because topics can be contrastive as well (cf. Krifka (2008) and Repp (2010)). Komen (2013:33) gives the following definition of focus:

(36) Focus is the part of the sentence that should be understood as most highlighted or salient by the addressee, because it is new with respect to the current mental model, or contrasts with presupposed information, or is unpredictable, non-recoverable or of high communicative interest.

This is a very intuitive and practical definition capturing a wide variety of possibilities, but it still contains some gradient notions that remain undefined. What exactly is unpredictable or when exactly is something of 'high' communicative interest? Krifka (2008) has furthermore shown that there does not need to be a correlation between given or well-established information (getting a linked label IDENTITY, INFERRED or NEW) and the distribution of focus: even well-established phrases with an IDENTITY label like pronouns can be focussed:

(37) Mary only saw [HIM]. (Krifka, 2008:39)

The capital letters in example (37) denote a stressed, rising accent and thus a focus on the pronoun. This example is perfectly fine in English, even though the referential state of the focussed pronoun is IDENTITY and thus linked or 'given'. In semantics, a constituent that is selected from a set of alternatives is understood to

be focussed (cf. Rooth (1985) and Zimmermann and Féry (2010)). Krifka's (2008) exact definition is as follows:

(38) A property F of an expression α is a Focus property iff F signals
  (a) that alternatives of (parts of) the expression α or
  (b) alternatives of the denotation of (parts of) α
  are relevant for the interpretation of α.

As long as 'relevant' is not further defined, this too leaves some room for subjective interpretation. If we want to investigate the information structure of a language we should not be distracted by possible phonological, morphological or syntactic *expressions* of IS. A high pitch accent, for example, may be used to focus a constituent in one language, but it does not necessarily have the exact same effect in another language. Nevertheless, there are certainly some cross-linguistic generalisations on the way IS is expressed. Ideally, we try to go *beyond* the surface expression to find its IS status first before we make the association between, e.g. high pitch and contrastive focus, or fronted constituents and topicalisation. Krifka's definition in (38) allows the separation of the way IS is expressed from what the IS status (referential state, focus domain, etc.) is. I therefore use the definition in (38) as a guideline to recognise focus constructions, or, in particular the domain that I generally label CONSTITUENT FOCUS. CONSTITUENT FOCUS can be marked in various ways, just like the topicalisation structures we noted above (see the sections on different types of focus and their markings below). Again, however, I can only discuss those forms of focus marking that can be detected in historical, written documents. Birch and Clifton (1995) showed in their experiments with *it*-clefts and *there*-insertions that structural positions can also make focus stand out in sentence comprehension tasks.

From a cognitive perspective, constituent focus structures play an important role in directing attentional focus in our brains. They also influence the availability of information in our memory and the degree to which it continues to be activated (Cowles, 2012:298). From psycholinguistic experiments we know that auditory cues like the pitch accents mentioned above can be helpful to identify focussed constituents (Cutler & Fodor, 1979). There is no consensus yet about a one-to-one mapping between prosody and information status (cf. Cowles (2012:293) and Hedberg and Sosa (2007)), but there is further evidence of these focussed structures from ERP studies. In some of those experiments, for example, N400 effects were detected when participants heard sentences with focus-violations (cf. K. Johnson (2003) for English and Hruska, Alter, Steinhauer, and Steube (2000) for German). The N400 effect, consisting of a characteristic change in brain wave activity 400 milliseconds after the stimulus, is associated with lexical and semantic processing (Kutas, Van Petten, & Kluender, 2006). This effect suggests focus anomalies influence the semantic processing of the word. Later studies on reading tasks with focus constructions by Bornkessel, Schlesewsky, and Friederici (2003), however, suggested that focus modulates information integration, indexed by a late positivity effect, instead of the N400 (cf. Cowles (2012)). Whichever it

turns out to be, it is clear that reading or hearing a focussed constituent results in a measurable effect in our brain. Although more experimental research is needed, a different focus domain, like THETIC FOCUS (see section 3.3.2) or PREDICATE FOCUS (see section 3.3.3) where the whole predicate is focussed instead of just one constituent, clearly gives the listener or the reader very different options.

Types of Constituent Focus

In section 3.3.3 on finding the right focus domain, we went through several steps to detect THETIC FOCUS and PREDICATE FOCUS. CONSTITUENT FOCUS, or 'narrow' or 'identificational' focus, as it is also called (cf. É.Kiss (1998)) can be found when there are alternatives of a certain expression that are relevant for the interpretation of the particular clause (see definition of Focus by Krifka (2008) above). Figure 3.2 shows the three focus domains, including the subtypes that can be detected in the domain of CONSTITUENT FOCUS:

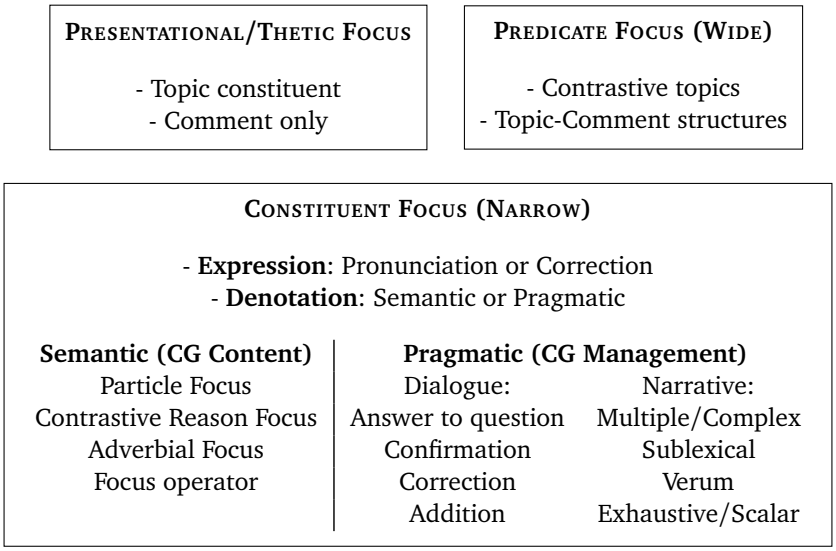


Figure 3.2: Focus Domains with subtypes

When we find relevant alternatives in a dialogue, we proceed to find out if the constituent is part of a question or answer. If it is not, we try and detect whether a constituent (or even part of it, a sublexical item) functions as a confirmation, correction or parallel structure. If this is the case, the clause will get the label of CONSTITUENT FOCUS with an addition: CONFIRMATION, PARALLEL and CORRECTION or another form of CONTRASTIVE FOCUS. If not, we are simply dealing with a topic-comment structure and thus label it PREDICATE FOCUS. (39) shows the schematic procedure just described. Examples (following Krifka's examples, unless indicated otherwise) of these types of focus are given in (40), (41), (42) and (43):

(39) Is it a question-answer dialogue?
- (i) Yes ⇝ Go to (45)
- (ii) No, did the speaker confirm information?
  - (i) Yes ⇝ CONFIRMATION FOCUS
  - (ii) No, did the speaker correct information?
    - (i) Yes ⇝ CORRECTION FOCUS
    - (ii) No, did the speaker use parallel structures?
      - (i) Yes ⇝ PARALLEL FOCUS
      - (ii) No, is there an explicit contrast?
        - (i) Yes ⇝ CONTRASTIVE FOCUS
        - (ii) No ⇝ PREDICATE FOCUS

(40) CONFIRMATION FOCUS

A: Siriol ate the last biscuit.
B: Yes, [SIRIOL] ate the last biscuit.

(41) CORRECTION FOCUS

A: Siriol ate the chocolate.
B: No, [ASIYE] ate the chocolate.

A: Theofiel?!
B: *Nee, Theo[DOOR] is mijn naam.*
No Theodoor is my name
'No, Theo[DOOR] is my name!'
(Dutch, from *De Texasrakkers*, Suske & Wiske 124)

(42) PARALLEL FOCUS

A DUTCH football fan talked to a ENGlish football fan about the world cup.

(43) CONTRASTIVE FOCUS

Martha: Woah, Nelly! I know for a fact you've got a wife in the country.
Shakespeare: But Martha, this is [TOWN].
The Doctor: Come on! We can have a good flirt later.
Shakespeare: Ooo, is that a promise, Doctor? *[winking at him]*
The Doctor: Oh, [FIFty-seven academics] just punched the air!
(from *Doctor Who*, series 3, episode 2)

The contrastive focus can be an explicit antonym or an alternative from a restricted set, as in the example above where *country* and *town* are contrasted. The contrast can also be implicit. The *fifty-seven academics* further on, for example, are raising their fists in victory, because they were just proven right: the phrase implies a contrast with all the other English literary scholars who do not think that Shakespeare was bi- or homosexual (referring to sonnet 57, which is about a relationship with a young man). If knowledge of English literary history is part of the world knowledge stored in the long-term memory of the reader, this contrast is obvious. Another example of implicit contrast is found in the following dialogue between someone hosting a workshop at a conference in Sydney and HRH the Earl of Wessex:

(44) A: May I invite you to join us for drinks, Sir?
B: Yes, why not? [In SYDney], I can safely go out.

The contrast in this utterance is obvious to those who know the British royal family and have dealt with their protocols before. In the UK, the Earl could never accept an invitation to go for drinks, because people will recognise him. In Australia, however, this is not the case.

Focus in dialogue
If we *are* dealing with a question-answer dialogue, we need to investigate the type of question: is it a wh-question and if so, does it extend over the entire VP or not? If it is not a wh-question, several other options remain: parallel answers (similar to parallel focus sentences above), delimitation focus and closed or open set answers. Consider the following continuation of the decision tree and the examples (after Krifka (2008), unless indicate otherwise):

(45) Is there a delimitation?
(i) Yes ⇝ DELIMITATION FOCUS
(ii) No, is it a simple wh-question?
(i) Yes ⇝ Go to (46)
(ii) No, is there a parallel answer?
(i) Yes ⇝ PARALLEL ANSWER
(ii) No, go to (46).

(46) Does focus extend over the entire VP or a NP/PP?
(i) Entire VP ⇝ VP WH-ANSWER
(ii) NP or PP, is it a closed or open set?
(i) Closed ⇝ CLOSED NARROW FOCUS
(ii) Open ⇝ OPEN NARROW FOCUS

(47) VP WH-ANSWER
A: What is Rhys doing?
B: He is [climbing Snowdon].

(48) PARALLEL ANSWER
A: Who ate what?
B: SIriol ate the BIScuit and ASiye ate the CHOcolate.

(49) DELIMITATION FOCUS
Which sister loves what?
a. As for ASiye, she loves CHOcolate.
Who do YOU think stole the chocolate?
b. In MY opinion, ASiye stole the chocolate.

(50) OPEN NARROW FOCUS
A: What would you like to drink?
B: I'd like some TEA, please.

A: Who is climbing Snowdon?
B: RHYS is climbing Snowdon.

A: How do you tell the story of pain?
B: You don't: you tell the story [of how], after everything falls apart, [you slowly rebuild].
(after `http://itellstories.com`, d.d. 31-12-12, *Twentytwelve*)

(51) CLOSED NARROW FOCUS
A: What would you like to drink, tea or coffee?
B: I'd like TEA, please.

Expression vs. Denotation Focus
If the clause under investigation is not part of a dialogue, the next question we ask is whether we are dealing with expression or denotation focus (cf. Krifka (2008:19-20)). Expression focus affects aspects like the choice of words or pronunciation; they do not have to involve meaningful units like constituents. When it affects the pronunciation, I label it PRONUNCIATION FOCUS. Another example of expression focus is found in corrections, e.g.:

(52) EXPRESSION FOCUS
Grandpa didn't [kick the BUcket], he [passed aWAY].

(53) PRONUNCIATION FOCUS
A: They live in BERlin.
B: They live in BerLIN.

Denotation focus is the most common form of focus outside dialogue situations. The first question here is whether we are dealing with semantic or pragmatic focus. According to Krifka (2008), pragmatic focus does not immediately influence truth conditions, but semantic focus *does* affect the truth-conditional content of the Common Ground. Contrastive focus is one of the best-studied cases of this type of focus. Semantic focus constructions are often clearly marked by semantic operators, such as focus-sensitive particles or adverbs like English *only*, *even*, *also* or *fortunately*, but this is not necessarily the case. The annotation procedure continues with the following decision-making tree:

(54) Is there an explicit lexical item as a semantic operator?
(i) No, go to (58).
(ii) Yes, are there more focussed constituents?
(i) Yes, go to (55).
(i) No, is there an adverbial focus operator?
(i) Yes ⇝ ADVERBIAL FOCUS
(ii) No, is it a negation or a particle?
(i) Negation ⇝ NEGATION FOCUS
(ii) Particle ⇝ PARTICLE FOCUS

(55) Are there two expressions introducing two different sets of alternatives?
(i) Yes ⇝ MULTIPLE FOCUS
(ii)No ⇝ COMPLEX FOCUS

Consider the following examples with more than one focussed constituent in (56) and (57) (from Krifka (2008:31-32)):

(56) MULTIPLE FOCUS
John only introduced BILL only to SUE.

(57) COMPLEX FOCUS
John only introduced BILL to SUE.

Example (56) contains two expressions introducing alternatives that are exploited in two different ways. The first *only* has scope over the second, reflected by a stronger accent on *Bill* than on *Sue*. This is not the case in (57) that only has one single focus on the pair <Bill, Sue>. If there is no overt semantic operator, we continue with (58):

(58) Is there a contrast with something in the CG?
(i) Yes ⇝ CONTRASTIVE FOCUS
(ii)No, is there a reason clause or variation of counterfactual?
(i) Yes ⇝ REASON CLAUSE FOCUS
(ii) No, start over (see Appendix for full procedure)

Krifka (2008) mentions (59) as an example of focus that I label REASON CLAUSE FOCUS:

(59) REASON CLAUSE FOCUS
a. **Clyde had to marry [BERtha]** in order to be eligible.
b. **Clyde had to [MARry]** Bertha for the inheritance.

Examples of CONTRASTIVE FOCUS can be found in many constructions and many different languages. Just like in the dialogue examples above, the contrast can be made explicit by repeating the same lexical item with a different modification (see (60) and (61)) or by using its antonym (or a close resemblance, see (63) and (62)). But it can also be implicit, contrasting the expected meaning of the items (as in (64)):

(60) The average pencil is [seven inches] long, with just a [half-inch] eraser, in case you thought optimism was dead. (Robert Brault)

(61) *Sans toi, les [émotions d' aujourd'hui] ne seraient que la peau morte*
without you the emotions of today NEG would ONLY the skin dead
des [émotions d' autrefois]
of.the emotions of past
'Without you, today's emotions would only be the dead skin of the emotions of the past.' (French, from *Amélie*)

(62) It is not enough for us to *believe* that what we do makes a difference - we must *prove* that it does, and be accountable to everyone we serve.
(from *Measuring the Award's impact*, B. Hirt (2012))

(63) *Wir vermögen [mehr], als wir glauben. Wenn wir das erleben, werden wir*
we can.do more than we think when we that realise will we
uns nicht mehr mit [weniger] zufrieden geben.
us not more with less satisfied give
'We are all better than we think. If (only) we can be brought to realise this we will never again be prepared to settle for anything less.'
(German, from Kurt Hahn)

(64) That's the whole problem with science. You've got a bunch of [empiricists] trying to [describe things of unimaginable wonder].
(from Calvin & Hobbes)

(65) When I meet you, in that moment, I'm no longer a part of [your future]. I start quickly becoming part of [your past]. But in that instant, I get to share [your present]. And YOU, you get to share MINE. And that is the greatest present of all. (from *Hiroshima* by Sarah Kay)

There is a wide variety of semantic operators that can indicate focus structures in different languages. Contrast can also play a role here, depending on the type of particle. Consider the following examples in Present-Day English and Welsh:

(66) PARTICLE FOCUS

a. *Dim ond gofyn am fenthyg sgriwdreifar ro'n i, nid adrodd hanes fy*
only ask.INF about borrow.INF screwdriver was i not relate story my
mywyd.
life
'I was only asking to borrow a screwdriver, not to relate the story of my life.'

b. *Dydyn nhw ddim yn gwneud dim byd eu hunain, dim ond dwyn*
are they NEG PROGR do.INF nothing themselves only steal.INF
oddi wrth eraill maen nhw.
from others are they
'They don't do anything themselves, they only steal from others.'
(from *Y rhyfel oeraf*, Baxendale (2009:43 and 89))

c. One of the great things about going to high school with people from 60 different countries was that we were all forced to see things, *even* the small, everyday things we all took for granted, from different perspectives.

d. I sincerely hope the results of our impact research framework will *not just* prove the value of this remarkable youth achievement award, but *also* convey the emotional effect.
(HRH The Earl of Wessex KG GCVO in *Measuring the Award's impact*, B. Hirt (2012))

Finally, there are some other types of focus we have not discussed yet. One further

question we can ask concerns the size of the constituent: is the entire constituent focussed or just part of it? Note that according to É. Kiss (1998), 'Identification Focus' (our CONSTITUENT FOCUS) can be distinguished from 'Information Focus' (the 'new information' often found in the topic-comment structures that I labelled PREDICATE FOCUS above) by the fact that only the latter can be smaller or larger than an XP as in (67):

(67) SUBLEXICAL FOCUS
Let me exPLAIN, exPOUND, exPAND and exPOSIT.
(from *A discussion on Language* in BBC's 'A bit of Fry & Laurie')

Strictly speaking, SUBLEXICAL FOCUS (see example (67)) cannot be part of 'Identification Focus' in her system. If we want to equate 'Identification' and 'Constituent' Focus domains, É. Kiss's categorie of 'Identification Focus' should be slightly expanded to ensure that it can capture every form of focus. Krifka (2008) furthermore mentions an extreme focus on the truth value of a sentence, VERUM FOCUS (see example (68) after Krifka (2008)).

(68) VERUM FOCUS
Asiye DOES like chocolate, why do you think she wouldn't?

There are furthermore two types of contrastive focus that we have not discussed: EXHAUSTIVE and SCALAR FOCUS (after Krifka (2008)):

(69) EXHAUSTIVE FOCUS
It's [ASIYE and ELANOR] that saved us.

(70) SCALAR FOCUS
Wild HORses wouldn't drag me there.

Example (69) is exhaustive in the sense that all possible candidates who could have 'saved us' were listed: Asiye and Elanor. Example (70) is scalar because it implies that there are more forces that could possibly 'drag me there', but even animals as strong as wild horses would not be able to do so (because I have made up my mind and really don't want to go). These last examples conclude a long section about many different types of CONSTITUENT FOCUS. In the next section, I discuss some ways to *mark* these focus structures.

Marking Constituent Focus

Evidence of CONSTITUENT FOCUS in historical data first of all comes from detecting possible alternatives relevant for the context. Once these possible alternatives have been found, we need to describe how they can be marked. As we have seen in topic marking above, in historical data we can only work with morphology, word order patterns, lexical items and, possibly, underlying syntactic structure. In the previous section, I already showed some examples of focus particles and other operators.

(71) **Focus Particles**

a. The leaves change colors in the fall. [People] change colors in the fall, **too**.
(from `http://itellstories.com`, d.d. 31-12-12 and 18-08-14)

b. *(...) y dywedir nad yw 'n rhewi* ***hyd yn oed*** *mewn*
... PRT said.IMPERS NEG.FOC is PROGR freeze.INF even in
gaeaf caled.
winter hard
'... it is said that it doesn't freeze, not even in a hard winter.'
(Modern Welsh)

c. *Does* ***dim ond*** *eisiau dechrau*
NEG.is only need begin.INF
'You only need to begin'
(Modern Welsh, from a poem by Ceiriog)

Special constructions like clefts are also commonly used in languages to mark focussed constituents:

(72) **Clefts, pseudoclefts and inverted pseudoclefts**

a. *Fi sydd ar fai am hynny.*
I is.REL on blame for that
'I am the one to blame for that.'
(Modern Welsh, Baxendale (2009:89))

b. *ma Se-rut hayta ze nexmada*
what that-Ruth was.F Z.M nice.F
'What Ruth was was nice.'
(Hebrew, Heller (1999:47))

c. There'll be days like this (...) when you step out of the phone booth and try to fly and the very people you want to save are the ones standing on your cape.
(from *Point B* by Sarah Kay via `www.kaysarahsera.com`)

Answers to questions furthermore often exhibit different word order patterns, depending on the type of question (yes/no, wh, broad/narrow focus, etc.):

(73) **Questions and answers**

a. *Wyt ti ffansi mynd am wibdaith fach 'te? Ydw, plis.*
are you fancy go for trip small TAG am please
'Do you fancy to go on a short trip then? I do, please.'
(Modern Welsh, Baxendale (2009:46))

b. *Pam mae 'r graig hon yn gynnes, tybed? Oherwydd nad craig*
why is the rock this PRED warm you-think because NEG.FOC rock
yw hi.
is it
'Why is this rock warm, you think? Because it is not a rock.'
(Modern Welsh, Baxendale (2009:92))

c. *Felly beth sy 'n digwydd nawr? Mae hi 'n amser mynd adref.*
So what is PROGR happen.INF now is it PRED time go home
'So what's happening now? It is time to go home.'
(Modern Welsh, Baxendale (2009))

In traditional grammars of Middle Welsh, focus structures are usually called 'mixed order': "[w]hen a part of the sentence other than the verb is to be emphasised, this is placed at the beginning of the sentence, preceded by a form of the copula and followed by a relative clause." (D. S. Evans, 2003 [1964]:140). Some examples he gives are (with his translation):

(74) **Mixed Order**

a. *Ys mi a 'e heirch.*
it-is me PRT her search.3S
'it is I who seek her' (Middle Welsh, WM 479.29)

b. *Oed maelgun a uelun i n imuan.*
was Maelgwn PRT saw.IPF.1S I PROGR fight.INF
'It was Maelgwn that I could see fighting.'
(Middle Welsh, YMTh 57.5)

In a later stage of the language, this sentence-initial copula was lost "before the emphasised word or phrase" (D. S. Evans, 2003 [1964]:141). Compare the following examples (again with Simon Evans's translation):

(75) **Mixed Order**

a. *Mi a 'e heirch.*
I PRT her search.3S
'(it is) I who ask for her' (Middle Welsh, WM 479.24)

b. *Mi yd wyt yn y geissaw.*
I PRT are.2S PROGR 3MS search
'(it is) I whom thou art seeking' (Middle Welsh, WM 138.21)

In these examples of the 'mixed order' there is no agreement between the subject and the verb. There is a very similar word order pattern in Middle Welsh, however, that does show agreement, but is not a focus structure:

(76) **Abnormal Order**

a. *Gwydyon a gerwys yn y blaen.*
Gwydyon PRT travelled.3SG in the front
'Gwydion travelled in the forefront' (not: 'It was Gwydyon who...')
(Middle Welsh, PKM 90.27)

b. *Mi a wn dy hanuot o 'm gvaet.*
I PRT know.1S 2S be.INF from 1S blood
'I know you are from my blood.' (Middle Welsh, CO 167)

This 'abnormal order' is often referred to as a topicalisation device (cf. Poppe (1991) and Willis (1998) among others). The first slot in this 'verb-second' construction can be filled by the subject, object or adjunct phrase (as we have seen in example (35) above). Finding the information-structural and syntactic constraints of these various word order patterns and how they change is the main research question of the present thesis. Chapter 4 presents a detailed description of IS in different stages of the Welsh language. The syntactic analysis of the various word order patterns in Chapter 5 sheds more light on the interface issues. For now it suffices to say that word order and syntactic relations interact with information structure in Welsh, so those above-mentioned markings of focus (and topic) structures will be investigated in more detail.

3.3.5 Focus domains of copula clauses

The three focus domains discussed above can also be found in copular clauses. Since the syntactic structure of copular clauses differs, I discuss the procedure of detecting the focus domains of these clauses separately. Komen (2013:164-170) gives a detailed overview of focus domains in copular clauses in English. In this section I propose a similar way of deriving the focus domain of copular clauses in Welsh, combining the coded syntactic and IS information, in particular the referential state of the core arguments. The focus domain is derived via a number of questions in a decision-making tree:

(77) Is it an equative clause?
  (i) Yes, move on to (79)
  (ii) No, is the subject NEW?
    (i) Yes ⇝ CONSTITUENT FOCUS as in (78a)
    (ii) No ⇝ PREDICATE FOCUS as in (78b)

(78) a. *Y mae Arthur yn gefnder iti.*
PRT be.PRES.3S Arthur PRED cousin to.2S
'Arthur is a cousin of yours.' (CONSTITUENT FOCUS - Modern Welsh)
  b. *Cauall oed y enw.*
Cafall be.PAST.3S 3MS name
'His name was Cafall.' (PREDICATE FOCUS - Gereint 399)

(79) Is the equative NP complement an Adjectival Phrase?
  (i) No, move on to (81)
  (ii) Yes, is the subject NEW?
    (i) No ⇝ PREDICATE FOCUS as in (80a)
    (ii) Yes ⇝ THETIC FOCUS as in (80b)

(80) a. *Roedd pawb yn 'gwybod' mai Jyrman Sbei oedd hi.*
was all PROGR know.INF that German spy was she
'Everyone knew that she was a German spy.'
(PREDICATE FOCUS - Modern Welsh)
  b. The world is wonderful. (THETIC FOCUS)

(81) Is the equative NP complement INERT?
(i) No, move on to (83)
(ii) Yes, Is the subject NEW?
(i) No ⇝ PREDICATE FOCUS as in (82a)
(ii) Yes ⇝ THETIC FOCUS as in (82b)

(82) a. *Ac Ioseph ydoedd fab deng mlwydd ar hugain pan...*
and Joseph be.PAST.3S lad ten year on 20 when...
'And Joseph was 30 when...' (PREDICATE FOCUS b1588 - Gen. 41.46)
b. In the next year Marius was consul. (THETIC FOCUS - Komen (2013:166))

(83) Is it a case of variable identification?
(i) Yes ⇝ CONSTITUENT FOCUS as in (84)
(ii) No, is the subject NEW?
(i) Yes ⇝ THETIC FOCUS as in (85)
(ii) No, is the subject INFERRED or ASSUMED?
(i) Yes, move on to (87)
(ii) No, is the subject INERT?
(i) Yes ⇝ PREDICATE FOCUS as in (86)
(ii) No, go to (87)

(84) CONSTITUENT FOCUS
a. *Y TARDIS yw hwn.*
the TARDIS is that
'That is the TARDIS.' (answer to: 'What's that?') (Baxendale, 2009:46)
b. (Last week, part of the Pont Des Arts in Paris collapsed. It collapsed, quite literally, under the weight of aspirations and expectations of everlasting love;) the Pont Des Arts was one of the famous bridges upon which young lovers would affix locks to signify the foreverness of their affection.
(from `http://itellstories.com`, d.d. 18-06-14, *Love locks*)

(85) *Maxen Wledic oed amherawdyr yn Ruuein*
Maxen Wledig be.PAST.3S emperor in Rome
'Maxen Wledig was emperor in Rome.' (THETIC FOCUS - BM 1.1)

(86) What is the weather in Siberia? In the winter, it is cold.
(PREDICATE FOCUS - Komen (2013:166))

(87) Is the complement NEW?
(i) Yes ⇝ CONSTITUENT FOCUS as in (88a)
(i) No ⇝ PREDICATE FOCUS as in (88b)

(88) a. *Gwidonot Kaer Loyw ynt.*
witches Gloucester be.3P
'They are the witches of Gloucester.' (CONSTITUENT FOCUS - Peredur 29.18-19)
b. The driver of that car is from Finland.
(PREDICATE FOCUS - Komen (2013:165))

3.3.6 Additional IS factors

As mentioned above, there are at least two further information-structural factors that can interact with each of the three focus domains: delimitation strategies or frame setters (see section 3.3.4 above) and the 'principle of natural information flow'. For every sentence we can detect one of the three focus domains, but we should further annotate these two notions to provide a comprehensive description of all IS facts.

Delimitation and Point of Departure

> *"**When you've told your love what you're thinking of***
> *things will be much more informal;*
> ***Through a sunlit land** we'll go hand-in-hand,*
> *drifting gently back to normal.*
> *(...)*
> ***With your hand in mine,** idly we'll recline*
> *amid bowers of neuroses,*
> ***While the sun seeks rest in the great red west***
> *we will sit and match psychoses".*
>
> (fragment from *The Passionate Freudian* by Dorothy Parker)

Delimitation strategies or 'points of departure' like the bold-faced phrases in the above poem by Dorothy Parker were already discussed in the section on topics (see section 3.3.3), because they are also called 'frame setting topics' (cf. Götze et al. (2007)). Krifka (2008) uses the term 'delimitation' for any expression (both frame setters and contrastive topics) that "always comes with a focus" generating alternatives (Krifka, 2008:48). This definition allows for more than just frame setters, e.g. (from Krifka (2008:48)):

(89) [An [inGEnious] mathematician]$_{\mathrm{Delim}}$ he is [NOT]$_{\mathrm{Focus}}$.

Komen (2013:44) gives the following definition of what they call 'Point of Departure' (PoD):

(90) **Point of Departure**
A point of departure is a constituent fulfilling the following conditions:
i) It is placed at the beginning of a clause or sentence;
ii) It expresses a change in the point of view in the discourse;
iii) It anchors to something that is accessible to the addressee (either from the preceding linguistic context or through shared knowledge)

I will label constituents that meet the requirements in (90) POINT OF DEPARTURE, because their presence can influence the IS status of the entire sentence. A sentence without a PoD is not as tightly linked to the previous context or content of the current Common Ground as sentences *with* a PoD. These types of frame setters occur very often in Middle and Early Modern Welsh (cf. Poppe (1991) where it is

called 'Situationskulisse'). To ensure all possible IS variables are covered, I make a further distinction between the functions of the PoDs. In this way if we encounter word order variation in different sentences, we could determine whether or not this is due to the different function of the PoD. Consider some examples of sentences with different PoDs below:

(91) PoD: LOCATIONAL

a. (I cycled to the office in the morning and worked all day.) **From the office**, I went straight to BodyCombat training.

(92) PoD: TEMPORAL

a. ***Et quand tu seras consolé*** *(...), tu seras content de m' avoir connu.*
and when you will.be consoled (...) you will.be happy of me have known
'And when you'll be comforted (...), you will be happy to have known me.'
(French, from *Le petit prince* by De Saint-Exupéry)

b. ***Om half 10*** *begint de handbalwedstrijd.*
at half 10 starts the handball game
'At half past nine, the game will start.' (Dutch)

(93) PoD: CIRCUMSTATIAL

a. **With an incredible amount of effort**, he managed to convince her.
b. **Healthwise**, my friend is fine.

(94) PoD: SITUATIONAL
As they had been friends for a long time, he expected her to help him.

(95) PoD: REFERENTIAL
That battery, however, continued its fire.

All of the above sentence-initial 'points of departure' contain information stored in the current CG: they all either refer back to something that was mentioned in the text or that is accessible as 'world knowledge' from our long-term memory. They set the frame or limit the space in which the following proposition holds. They can be added to clauses with any of the three focus domains: THETIC FOCUS, PREDICATE FOCUS or CONSTITUENT FOCUS.

Principle of Natural Information Flow

Another IS phenomenon that can interact with each of the three focus domains is what Comrie (1989), Kaiser and Trueswell (2004) and others have called the "Principle of natural information flow" (cf. Komen (2013:43-44)). This principle concerns the degree of 'givenness' of constituents: established information precedes less established information. If the syntactic structure of the language allows for alternatives, some constituents can be reordered changing the 'information flow' of the sentence. We can see the principle in presentational constructions in English (cf. Komen (2013:44)):

(96) UNMARKED INFORMATION FLOW
Once upon a time there was **a handsome prince**.

The referential state of the phrase *a handsome prince* is NEW and it is thus placed at the very end of the sentence. In the English Dative Alternation we also see a clear example of this principle:

(97) UNMARKED INFORMATION FLOW
a. Rhys gave the student **a book**.
b. Rhys gave the book to **a student**.

Both examples in (97) abide by the principle of information flow, because in both cases (as the definite article shows), the first constituent following the verb conveys 'more established' information than the second constituent. Note that the opposite word order in English with the same noun phrases is odd or even impossible:

(98) MARKED INFORMATION FLOW
a. Rhys gave a book to the student.
b. Rhys gave a student the book.

In some constructions in English, however, putting the least-established constituent before the rest has a special effect, for example, to focus the place in the Locative Inversion or the direct object that has been the centre of attention of the entire lecture, as in example (99a) and (99b):

(99) MARKED INFORMATION FLOW
a. **Up, up, up the stairs** we go!
(from *The Lord of the Rings* by JRR Tolkien)
b. Sir William Jones and John James Jones both worked tirelessly to bring to a world far distant in time and place **some of the wealth of ancient Indian culture**.
(from a lecture on JJ Jones and the *Mahavastu* by Silk (2014:439))

The Principle of Natural Information Flow can occur with any of the three focus domains. All clauses are annotated as MARKED (unlinked before linked) or UNMARKED (linked before unlinked) for this in the Welsh database.

3.4 Conclusion

In this chapter I gave an introduction to Information Structure and its place in the field of linguistics. I discussed three core information-structural notions in greater detail: Givenness, Topic (vs. Comment) and Focus (vs. Background). For each of these notions, I outlined their main characteristics in a systematic way so that they can be used to annotate a corpus consistently.

For the notion of Givenness, it is clear that a simple binary distinction between Old and New information is not enough (see Taylor and Pintzuk (2014) for a

systematic evaluation of different annotation schemes). For the present thesis, I annotated the referential status of subjects and objects in the Middle Welsh corpus according to the Pentaset developed by Komen (2013). This type of annotation can help identify effects in word order distributions in combination with annotated syntactic features.

In the section on Topics, I focussed on three different kinds of topics that are found in the Middle Welsh corpus: aboutness, contrastive and familiar topics. The notion of 'Delimitation' as formulated by Krifka plays a crucial role in determining aboutness topics. Like frame or scene setters, they usually occupy the first position in the sentence. Contrastive topics are also found in Middle Welsh. The notion of contrast is thus not necessarily associated with Focus. In final part of this thesis, these kinds of topics are discussed again in their syntactic contexts.

I furthermore presented a detailed overview of different kinds of Focus structures. I illustrated the different types observed in the literature with examples from Welsh and various other languages. I furthermore presented some systematic 'algorithms' to find the focus articulation of copular clauses, based on studies in the history of English by Komen (2013).

Finally, I discussed two further notions that are relevant to information structure: Point of departure and Information Flow. Many so-called 'Points of Departure' of a sentence appear in the form of temporal or circumstantial clauses. In effect, they function as frame setters delimiting the context of the rest of the sentence. The Principle of Natural information flow finally stipulates that old information usually precedes new information. In sentences with the reverse order, the 'flow' of information, or in particular the referential status of the core arguments, is 'marked'.

These three core notions of Givenness, Topic and Focus, in combination with the additional annotation for specific points of departure and information flow are argued to provide a comprehensive insight into the Information Structure of the sentence in its context. The clear definitions and guidelines to find the right labels presented in this chapter facilitate annotation. A consistent analysis of this kind helps to make the study of Information Structure that has suffered from a lot of 'terminological profusion and confusion' more insightful in the language under investigation. But, more importantly, it renders it more useful, because results of such thorough investigation could then be more easily compared between different languages.