Abstract

This paper, in the context of multilingual MT, proposes the use of ICONS (Individual CONstraintS) to add a representation of information structure to MRS. The value of ICONS is a list of objects of type *info-str*, each of which has the features CLAUSE and TARGET. The subtypes of *info-str* indicate which information structural role is played by the TARGET with respect to the CLAUSE. This proposal is designed to support both the calculation of focus projection from underspecified representations and the handling of multiclausal sentences.

1 Introduction

This paper presents an HPSG (Pollard and Sag, 1994) analysis of information structure marking, with an eye towards practical applications such as machine translation (MT), adding constraints on information structure to MRS (Copestake et al., 2005) representations. In particular, we aim to improve on our previous analysis presented in Song and Bender (2011), to overcome two difficulties facing that work: First, we did not specify how the analysis could handle the spreading of focus beyond the lexical item directly marked for focus. Second, by encoding information structure as constraints on features of semantic variables ('variable properties'), we predicted that all occurrences of an index could share the same information structural properties. This is not necessarily the case, especially in constructions where semantic indices are shared across multiple clauses. This paper suggests the use of individual constraints (henceforth, ICONS), which (i) leave the information structural values of some constituents underspecified, facilitating an analysis of focus projection, and (ii) allow us to anchor the constraints on information structure with respect to the clause they belong to.

This study aims to provide a theoretical framework to create a grammar library for information structure, which will be added to the LINGO Grammar Matrix

[†]First of all, we are especially grateful to Dan Flickinger and Ann Copestake for the idea of using ICONS for information structure. Thanks also to Woodley Packard for adding support to ICONS to the ACE generator (`http://sweaglesw.org/linguistics/ace`)), which allowed us to confirm the feasibility of our proposal. Russian and Japanese judgments reported in this paper were provided by Varya Gracheva, Zina Pozen, and Sanae Sato. We also thank Frank Van Eynde, Berthold Crysmann, Kiyong Lee, Yo Sato, and David Erschler for their comments and suggestions at the venue, and three anonymous reviewers for helpful feedback. After the conference, the first author had several opportunities to discuss some parts of our proposal with several linguists in Korea, which helped us refine our proposal once again. Though it should be noted that we could not fully accommodate their suggestions in this paper, we thank Jong-Bok Kim, Jae-Woong Choe, Hae-Kyung Wee, and Young Chul Jun. All remaining errors and infelicities are our own.

This material is based upon work partially supported by the National Science Foundation under Grant No. 0644097. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

(Bender et al., 2002, 2010).[1] The LINGO Grammar Matrix is an environment for developing precision grammars from a typological perspective. The Grammar Matrix customization system, in particular, functions as a starter-kit for the creation of computational grammars within the HPSG and MRS framework. That means this study has to (i) refer to cross-linguistic findings about information structural meanings and forms to express information structure, and (ii) suggest a range of computational models described in the DELPH-IN joint reference formalism (TDL; Copestake 2002), which (ii-a) deal with different types of information structure marking found in the world's languages and (ii-b) constrain the MRS to reflect the information structure encoded by the marking.

This paper is structured as follows: §2 offers a brief explanation of information structural components and forms of expressing information structure in the languages this paper is concerned with. §3 proposes an analysis based on Individual Constraints. Comparing to previous studies, §4 and §5 show that our proposal processes information structure in a more effective way. Building on the analyses, §6 presents a sample translation from English to Japanese, and §7 explains how our proposal has been implemented and shows the outputs are as expected.

2 Information Structure

2.1 Components of Information Structure

This paper starts from the following assumptions, consistent with Song and Bender (2011): (i) Information structure consists of three components, namely, focus, topic, and contrast. (i-a) Every sentence presumably has at least one focus, while all sentences do not always have a topic. (i-b) Contrast, contra Lambrecht (1996), is treated as an information structural component in that it can be linguistically expressed. (i-c) Sometimes, a linguistic item can convey the meaning of neither focus nor topic, which we call background (a.k.a. tail, represented as *bg* in the hierarchy of this paper). (ii) Semantically empty and syncategorematic categories (e.g. expletives, semantically empty auxiliaries) are informatively empty as well; thus, they cannot signal any information structural meanings.

Focus refers to what is informatively new and/or important in the sentence (Lambrecht, 1996). This leads to an important linguistic property that distinguishes focus from other components: the focus of a sentence (as used in a particular context) can never be omitted, while topic and background elements can. *Wh*-questions have been employed as a tool to probe the focus meaning and marking: For instance, if the question is *What barks?*, the constituent corresponding to the *wh*-word in the answer bears focus. In English, this is typically marked with the the A-accent (H*), as in *The* DOG *barks.*[2]

[1] The LINGO Grammar Matrix has been developed in the context of the DELPH-IN consortium (`http://www.delph-in.net`).

[2] In this paper, SMALL CAPS stands for an A-accented phrase, **boldface** for a B-accented one, and $[_f\ ]$ for focus projection.

Topic is what an utterance is about. As mentioned in the previous paragraph, some languages (a.k.a. topic-drop languages (Huang, 1984)) frequently drop topics from sentences; thus, topics do not always appear overtly in running text or speech. Choi (1999) suggests the tell-me-about test for identifying topic: e.g. In a reply to *Tell me about the dog.* an NP referring to the dog will be the topic. In English, this can be marked with the B-accent (L+H*): *The* ***dog*** BARKS.

Contrast (realized as either contrastive topics or contrastive foci) always entails an alternative set, and can be expressed lexically (e.g. *thì* in Vietnamese (Nguyen, 2006)) or syntactically (e.g. preposing to the initial position in Standard Arabic (Ouhalla, 1999)), depending on the language. Several tests to detect contrast, such as the conditional test (Wee, 2001) for contrastive topic, the correction test (Gryllia, 2009) for contrastive focus, have been suggested, though they are not always cross-linguistically valid.[3]

2.2 Languages

While the analysis we develop is intended to be flexible enough to work cross-linguistically, we will use English, Japanese and Russian to exemplify three common types of information structure marking. English primarily uses prosody for this function (e.g. A/B-accents (Jackendoff, 1972)).[4] Japanese employs morphological markers: For instance, if the topic marker *wa* is attached to an NP, the NP involves either topic or contrast, or both (i.e. contrastive topic). On the other hand, if the case markers (e.g. *ga* for nominatives) are used instead of *wa*, the NP cannot fill the role of topic (Heycock, 1994). In contrast to English and Japanese, Russian takes advantage of its relatively free word order to assign a specific position to signal focus: Non-contrastive focus appears clause-finally and contrastive focus is preposed (Neeleman and Titov, 2009). The major patterns of expressing information structure in these languages are summarized are summarized in Table 1.[5]

2.3 Differences in Felicity

Information structure affects the felicity of a sentence in different discourse contexts. Sets of allosentences (i.e. close paraphrases which share truth conditions (Lambrecht, 1996)) differing only in information structure will differ in felicity in

[3] Hae-Kyung Wee and Young Chul Jun, p.c.

[4] There seems to be no consensus regarding this generalization. Dissenting views include Steedman (2000) based on a study of the interface between syntax and phonology and Hedberg and Sosa (2007) from the perspective experimental phonology, among others. Here we are not concerned with a precise account of the phonological realization of information structure marking in English, but rather how to represent the information structural effects of that marking for computational purposes. Therefore, we provisionally take Jackendoff's notion of A and B accents as a stand in for the prosodic representation.

[5] Of course, these languages can make use of others means to express information structure. English also has syntactic means to lay focus a constituent, such as clefts, pre-subject position, etc. So-called scrambling in Japanese also constrains information structure (Ishihara, 2001). Accents can also be used to signal focus in Russian.

Table 1: Languages

	English (eng)	**Japanese** [jpn]	**Russian** [rus]
means	prosody	lexical marking	syntactic positioning
focus	A-accent	case markers (*non-topic*)	clause-final
topic	B-accent	topic marker *wa*	unknown
contrast	A/B-accent	*wa*+scrambling	preposing (*contrast-focus*)

a given context. Multilingual NLP systems (e.g. MT) can be improved by making them sensitive to such constraints. For example, *The dog barks.* can be translated into at least two sentences in Japanese and Russian respectively. If *dog* bears the B-accent in English, the corresponding Japanese word *inu* should be combined with the topic marker *wa*, and the corresponding Russian word *sobaka* cannot occur clause-finally, as given in the first column of (1). On the other hand, if *dog* bears the A-accent, the nominative maker *ga* has to be used in Japanese, and the corresponding word can show up clause-finally in Russian, as shown in the second column of (1).[6]

(1) a. The **dog** BARKS. | The DOG barks.

b. inu-wa hoeru | inu-ga hoeru
dog-TOP bark dog-NOM bark [jpn]

c. sobaka laet | laet sobaka
dog bark bark dog [rus]

3 Individual Constraints

We propose to represent information structure via a feature ICONS (Individual CONstraintS) added to structures of type *mrs* (i.e. under CONT) as in (2). ICONS represents information structure as a binary relation between individuals and events. The items on the ICONS list are feature structures of type *info-str*[7] which indicate which index (the value of TARGET) has an information structural property and with respect to which clause (the value of CLAUSE). ICONS behaves analogously

[6] Angelina Ivanova and David Erschler each pointed out to us that the first sentence of (1c) can lay focus on *sobaka*, if the word bears a specific accent. That means the sentence *sobaka laet* could be ambiguously interpreted, if it were not for the accent on the subject *sobaka*. What we particularly argue is that the second sentence of (1c), in which the subject is overtly postposed, cannot correspond to the first sentences of (1a-b), because there is an obvious clue for the focus meaning whereby the sentence *laet sobaka* becomes unambiguous unlike the first sentence *sobaka laet*.

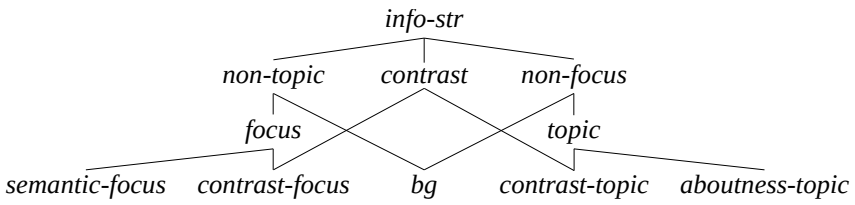
[7] The feature ICONS was originally proposed by Ann Copestake and Dan Flickinger, for the purpose of capturing semantically relevant connections between individuals which are nonetheless not well modeled as elementary predications, such as those found in intrasentential anaphora, apposition, and nonrestrictive relative clauses. Copestake and Flickinger recognized that we could use this same mechanism to anchor information structural constraints to particular clauses. In a more general system that uses ICONS both for our purposes and its original goals, the value of ICONS would be a list of items of type *icons*, where *info-str* is a subtype of *icons*.

to HCONS and RELS in that values of *info-str* are gathered up from daughters to mother up the tree.

(2)
$$\begin{bmatrix} \textit{mrs} & \\ \text{HOOK} & \begin{bmatrix} \textit{hook} & \\ \text{LTOP} & \textit{handle} \\ \text{INDEX} & \textit{individual} \\ \text{XARG} & \textit{individual} \\ \text{ICONS-KEY} & \textit{info-str} \\ \text{CLAUSE-KEY} & \textit{event} \end{bmatrix} \\ \text{RELS} & \textit{diff-list} \\ \text{HCONS} & \textit{diff-list} \\ \text{ICONS} & \left\langle ! \ldots, \begin{bmatrix} \textit{info-str} & \\ \text{CLAUSE} & \textit{individual} \\ \text{TARGET} & \textit{individual} \end{bmatrix}, \ldots ! \right\rangle \end{bmatrix}$$

In a particular ICONS element, the type will typically be resolved from *info-str* to a more specific type, drawn from the hierarchy in (3), to indicate the particular information structural role played by the TARGET in the CLAUSE. The *info-str* hierarchy is inspired by the analogous hierarchy from Song and Bender (2011), but is extended with three additional nodes: *non-topic*, *non-focus*, and *bg*: (i) *non-topic* means the target cannot be read as topic (e.g. case-marked NPs in Japanese); (ii) *non-focus* similarly indicates that the target cannot be the focus, and would be appropriate for e.g. dropped elements in pro-drop languages; (iii) finally, *bg* (background) means the constituent is neither *focus* nor *topic*, which typically does not involve additional marking but may be forced by particular positions in a sentence.

(3)



The type hierarchy (3) has three merits, comparing to our previous version presented in Song and Bender (2011) and other approaches in previous literature. First, (3) reveals that *contrast*, which is in a sister relation to *non-topic* and *non-focus*, behaves independently of *topic* and *focus* themselves. It has often been observed that a constituent in a language can convey an ambiguous meaning (i.e. contrastive meanings *vs.* non-contrastive meaning) even though it is marked in a specific form to express information structure in the language, and the meaning can be resolved only depending upon the given context in many cases. In order to represent the undetermined meanings properly in MRS, it is necessary to use a more flexible hierarchy which involves a cross-classification between *contrast*

and *topic/focus*. Second, *non-topic* and *non-focus* facilitate more flexible representation for informatively undetermined items in some languages. For example, case-marked NPs can convey the meaning either focus or background in Japanese (Heycock, 1994). That is, since a Japanese case marker (i.e. *ga* for nominatives) can convey two information structural meanings (*focus* and *bg*), the marker itself has to be less specifically represented as *non-topic* that both *focus* and *bg* inherit from. Third, we can make use of *bg* as a cross-cutting category, which sometimes needs to be explicitly marked. For instance, in English cleft constructions, the remaining part of the sentence after the relative pronoun should be represented as *bg*, because English cleft constructions belong to *focus-bg* in terms of sentential forms (Song and Bender, 2011).

The value of ICONS is constrained by both lexical and phrasal types. First, every lexical entry that introduces an index which can participate in information structure inherits from *icons-lex-item* (4a). This type bears the constraints which introduce an ICONS element as well as providing a pointer to the ICONS element inside the HOOK (ICONS-KEY), for further composition. *Icons-lex-item* also links the HOOK|INDEX to the TARGET value. On the other hand, lexical entries that cannot play a role in the information structure (e.g. semantically void lexical entries, such as case marking adpositions) inherit from *no-icons-lex-item* (4b), which provides an empty ICONS list.

(4) a.
$$\left[\begin{array}{l} \textit{icons-lex-item} \\ \text{HOOK}\left[\begin{array}{ll}\text{INDEX} & \boxed{1}\\ \text{ICONS-KEY} & \boxed{2}\end{array}\right] \\ \text{ICONS}\left\langle !\ \boxed{2}\left[\text{TARGET}\ \ \boxed{1}\right]!\right\rangle \end{array}\right]$$

b.
$$\left[\begin{array}{l} \textit{no-icons-lex-item} \\ \text{HOOK}\left[\begin{array}{ll}\text{ICONS-KEY}|\text{CLAUSE} & \boxed{1}\\ \text{CLAUSE-KEY} & \boxed{1}\end{array}\right] \\ \text{ICONS}\left\langle !\ !\right\rangle \end{array}\right]$$

Because the CLAUSE value needs to reflect the position in which a constituent is realized overtly, it is constrained via the phrase structure rules. Verbs which head their own clauses (i.e., finite verbs, plus certain uses of non-finite verbs) identify their CLAUSE value with their own INDEX (and thus their own TARGET) as shown in (5a).[8] For elements that do not head clauses, the CLAUSE value is constrained to be the INDEX of the verbal projection they attach to by *head-icons-phrase* (5b). This type is supertype to headed rules which can constrain information structure: e.g. *head-subj-phrase*, *head-comp-phrase*, and *head-mod-phrase*.

[8]The restriction to clause-heading verbs is meant to allow for examples like *The dog sitting on the mat barks.* where we believe that all elements of the VP *sitting on the mat* should take the INDEX of *barks* as their CLAUSE, not that of *sitting*.

(5) a.
$$\begin{bmatrix} \textit{verb-lex} & \\ \text{HOOK} & \begin{bmatrix} \text{INDEX} & \boxed{1} \\ \text{CLAUSE-KEY} & \boxed{1} \\ \text{ICONS-KEY|CLAUSE} & \boxed{1} \end{bmatrix} \end{bmatrix}$$

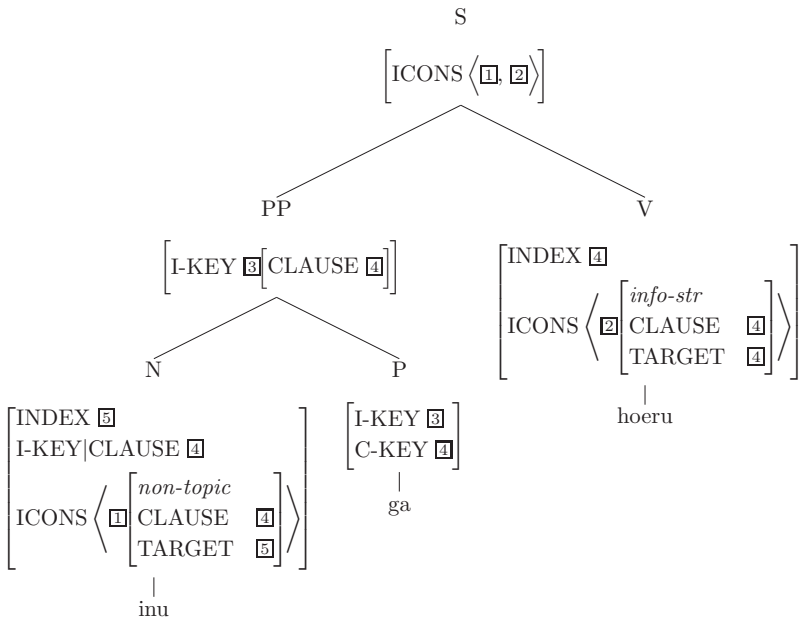
b.
$$\begin{bmatrix} \textit{head-icons-phrase} & \\ \text{HD-DTR|...|HOOK|CLAUSE-KEY} & \boxed{1} \\ \text{NON-HD-DTR|...|HOOK|ICONS-KEY|CLAUSE} & \boxed{1} \end{bmatrix}$$

The type of the ICONS-KEY value of a constituent (which, recall, points to an element of the ICONS list) can be constrained by accents responsible for information structural meanings, lexical rules attaching information structure marking morphemes, phrase structure rules corresponding to distinguished positions, or particles like Japanese *wa* combining as heads or modifiers with NPs. The headed rules can have subtypes which handle information structure differently, resolving the type of an ICONS element or leaving it underspecified. For example, the Russian allosentences (1c) are instances of *head-subj-phrase*, but the first one (*sobaka laet*), in which the subject is in situ, is licensed by a subtype that does not resolve the ICONS value, while the second one (*laet sobaka*), in which the subject is marked through being postposed, is licensed by the one which does. Hence, as shown in (7), the in-situ subject in Russian is specified as *info-str* (i.e. underspecified), whereas the overtly postposed subject is specified as *focus*.

The strategy of having phrase structure rules constrain the CLAUSE value of ICONS elements runs into a potential problem with *head-comp-phrase* because this rule is used in many different ways in our grammars. In particular, the problem arises with elements like Japanese case-marking adpositions: *inu-ga* 'dog-NOM' is an instance of *head-comp-phrase*, but *inu* has no informational structural relation with its head *ga*, and *ga* itself is semantically empty and thus has an empty ICONS list.[9] On the other hand, when *head-comp* joins a verb with its object (such as the PP *inu ga*), we want to connect the object's CLAUSE to the verb's INDEX. Rather than creating subtypes of *head-comp* to handle this differing behavior, we add the feature CLAUSE-KEY to mediate between the INDEX of the head and the CLAUSE value of the dependent. The phrase structure rules identify the head's CLAUSE-KEY with the non-head's ICONS-KEY|CLAUSE. Clause-heading verbs identify their INDEX and CLAUSE-KEY values. Case marking adpositions, on the other hand, inherit from *no-icons-lex-item*, which identifies CLAUSE-KEY with ICONS-KEY|CLAUSE. Note, however, that the value of ICONS-KEY is not identified with anything on the actual ICONS list for these elements, allowing ICONS-KEY|CLAUSE to function as sort of a scratch slot.

[9] On why *ga* et al are best treated as postpositions rather than affixes, see Siegel (1999).

(6) a.



b.

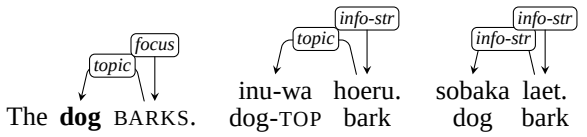
$$\begin{bmatrix} \textit{mrs} \\ \text{LTOP} \quad \boxed{h1} \\ \text{INDEX} \quad \boxed{e2} \\ \text{RELS} \quad \left\langle \begin{bmatrix} \textit{\_inu\_n\_rel} \\ \text{LBL} \quad \boxed{h3} \\ \text{ARG0} \quad \boxed{x4} \end{bmatrix}, \begin{bmatrix} \textit{exist\_q\_rel} \\ \text{LBL} \quad \boxed{h5} \\ \text{ARG0} \quad \boxed{x4} \\ \text{RSTR} \quad \boxed{h6} \\ \text{BODY} \quad \boxed{h7} \end{bmatrix}, \begin{bmatrix} \textit{\_hoeru\_v\_rel} \\ \text{LBL} \quad \boxed{h8} \\ \text{ARG0} \quad \boxed{e2} \\ \text{ARG1} \quad \boxed{x4} \end{bmatrix} \right\rangle \\ \text{HCONS} \quad \left\langle \begin{bmatrix} \textit{qeq} \\ \text{HARG} \quad \boxed{h6} \\ \text{LARG} \quad \boxed{h3} \end{bmatrix} \right\rangle \\ \text{ICONS} \quad \left\langle \begin{bmatrix} \textit{non-topic} \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x4} \end{bmatrix}, \begin{bmatrix} \textit{info-str} \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{e2} \end{bmatrix} \right\rangle \end{bmatrix}$$

Building upon the constraints presented so far, a sample derivation for a Japanese sentence is illustrated in (6a): First, CLAUSE-KEY of the nominative marker *ga* is identified with its own ICONS-KEY|CLAUSE. Second, when the *head-comp-phrase* combines *inu* and *ga*, the ICONS-KEY|CLAUSE of *inu* is identified with the CLAUSE-KEY of *ga*, in accordance with *head-icons-phrase*. The ICONS-KEY of *ga* is passed up to the mother (Semantic Inheritance Principle). When the *head-subj-phrase* combines *inu-ga* and *hoeru*, the ICONS-KEY|CLAUSE of the subject *inu-ga* (and thus of both *inu* and *ga*) is identified with the INDEX of *hoeru*. The corresponding MRS representation is given in (6b).

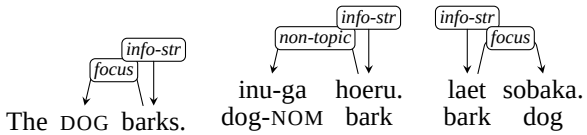
3.1 How MT works via ICONS

In the remainder of the paper, we will present information structural constraints in the style of dependency graphs of DMRS (Dependency MRS; Copestake 2009), for ease of exposition. In these graphs, the ICONS values are represented as links between head nouns (introducing the referential index that is the value of TARGET) and verbs (introducing the event variable that is the value of CLAUSE) and as unary properties of verbs themselves.[10] The graphs of the translations given in (1) are sketched in (7). The dependency graphs in (7) illustrate how our proposal gives rise to underspecified representations when information structure is not explicitly marked. Unless there is a specific clue to identify information structure such as A/B-accents in English, the topic marker *wa* in Japanese, and the clause-final position in Russian, the ICONS value remains just *info-str*.

(7) a.



b.



In (7a), the first graph represents an English sentence in which the subejct *the **dog*** bears the B-accent, thereby plays the role of topic, while the verb BARKS with the A-accent conveys the focus meaning. The direction of arrow stands for the binary relation between a TARGET (an entity) and a CLAUSE that the TARGET belongs to. The arc from BARKS to **dog** means the index of **dog** has a topic relation to the index of BARKS. The arrow to BARKS means the verb is linguistically marked as focus, with respect to the clause that it heads. The second graph in (7a) represents the Japanese translation, but since *hoeru* corresponding to BARKS has no overt mark of information structure, it remains just underspecified as *info-str*, differently from BARKS in the first graph. Likewise, in the third graph, since there is no information structural clue on *sobaka* corresponding to **dog** in the English translation, it also remains underspecified.

This ability to partially specify information structure allows us to reduce the range of outputs in translation while still capturing all legitimate possibilities. As mentioned in the foonote 6, the unmarked Russian sentence *sobaka laet* itself can correspond to both *The **dog*** BARKS in (7a) and *The* DOG *barks* in (7b), unless a phonological factor signals focus. That means, *The **dog*** BARKS can be translated into only *sobaka laet* corresponding to the third graph in (7a), but the same Russian sentence can be translated into both *The **dog*** BARKS and *The* DOG *barks*.

[10]This difference is because we use the event variable introduced by the verb to represent the clause, thus in the *info-str* constraint on the ICONS list of a verb, the TARGET and CLAUSE values are identified (cf. (5a)). Note also that though our examples focus on nominal arguments of verbs, the analysis is intended to scale to all semantically contentful elements.

3.2 Comparison to Previous Studies

The first main difference between our approach and previous studies has to do with the calculation of focus projection and in particular the role of underspecification. (8) provides a simple example of focus projection. The overt mark of focus is the A-accent on DOG, but this can be interpreted as spreading only to the NP or as spreading or projecting to the entire sentence. These different interpretations have different felicity conditions. The first could be the answer to the question *What barks?* (i.e. *focus-bg*), while the second to the question and *What happens?* (i.e. *all-focus*).

(8) a. [$_f$ The DOG] barks.

b. [$_f$ The DOG barks.]

Regarding the interpretation of (8), we can assume that (i) the two readings correspond to two distinct structures (parse trees), or (ii) the two readings are further specializations of one MRS, which is associated with one syntactic structure and includes some underspecified values. Here, as our goal is a computational model, we take the second approach for practical reasons and underspecify the type of the ICONS element for unmarked constituents such as *barks* in (8). Some previous work (Engdahl and Vallduví, 1996; De Kuthy, 2000; Chung et al., 2003), in contrast, takes the first approach without using underspecification: All sentences, within these frameworks, have as many syntactic trees as potential information structural interpretations.

Second, our approach has both similarities and differences to earlier work representing information structure in MRS. Wilcock (2005) models the scope of focus similarly to quantifier scope (i.e. HCONS), which is close to the idea that we take as our departure point for discussion. The difference between Wilcock's proposal and ours is that information structure in his model is represented as variables over handles, but ICONS captures the clause that an individual informatively belongs to as a binary relation, which facilitates scaling to multiclausal constructions.

(9) a. The president [$_f$ hates the china set].

b. 1:the(x,2), 2:president(x), 3:the(y,4), 4:china(y), 4:set(y), 5:hate(e,x,y)
TOP-HANDLE:5, LINK:1, FOCUS:3,5 (wide focus)

For instance, (9b) taken from Wilcock (2005, p. 275) represents the wide focus reading of (9a) (i.e. from 3 to 5). Note that in this representation, LINK (*topic* in this paper) and FOCUS have no relation to the clause or its head (*hate*). Paggio (2009) also models information structure within the MRS formalism, but information structural components in her proposal are represented as a part of the context, not the semantics. Though each component under CTXT|INFOSTR involves coindexation with individuals in MRS, her approach cannot be directly applied to the LOGON MT infrastructure that requires all transfer-related ingredients accessible in MRS (Oepen et al., 2007). Bildhauer and Cook (2010) offer an MRS-based architecture, too: Information structure in their proposal is represented directly under

SYNSEM (i.e. SYNSEM|IS) and each component (e.g. TOPIC, FOCUS) has a list of indices identified with ones that appear in EPs in SYNSEM|LOC|CONT|RELS, which is not applicable to the LOGON infrastructure for the same reason.[11]

In short, using ICONS has two merits in the context of implementing NLP systems; (i) underspecifiability, and (ii) the availability of a binary relation between individuals. The former facilitates flexible, partial representations and the latter enables us to capture information structure even in multiclausal sentences. The following sections cover each of these points in turn.

4 Underspecifiability

Previous approaches to the modeling of information structure are not efficient in NLP systems because having a large number of trees eventually has an adverse effect on performance as well as accuracy. Since it is important for transfer-based MT to reduce the number of potential analyses in each step, it is necessary to use a more effective and flexible method to represent information structure. We believe that our underspecified representations[12] can be further constrained to represent different information structural interpretations (consistent with the given ICONS list) in the same way that scope-underspecified MRSs can be further constrained with handle identities to yield fully scoped representations consistent with the given HCONS list. Thus, similarly to how a sentence with a scopal ambiguity (e.g. *Every dog chases some white cat.*) has a single MRS partially constrained via *qeq*s, the current work proposes that (8) be given a single representation whose information structure is partially constrained via ICONS.

We leave the development of the algorithm that calculates focus projection over MRS+ICONS to future work. We are particularly interested to investigate whether the MRS structure augmented with ICONS is sufficient, or if the focus projection algorithm would require access to syntactic structure. We note that previous work on focus projection (De Kuthy, 2000; Chung et al., 2003) highlights the importance of grammatical functions. However, the relevant distinctions (argument *vs.* adjunct status, peripheral *vs.* non-peripheral arguments) can be reconstructed on the basis of the MRS alone. Therefore, we consider it at least plausible that MRS+ICONS will contain enough information to calculate the range of fully-specified information structures for each sentence.

[11] We, of course, do not claim that every grammar should be compatible with the LOGON infrastructure. As mentioned in the introduction, the ultimate goals of this study include creating a computational library within the Grammar Matrix, which can be effectively used to enhance performance of HPSG/MRS-based MT systems. Given that LOGON, for now, is the readily available infrastructure for the purpose, our approach follows the requirements as far as possible.

[12] In early work on information structure in HPSG, Kuhn (1996) also suggests an underspecified representation for information structure, noting that prosodic marking of information structure often yields ambiguous meanings, which cannot in general be resolved in computational sentence-based processing.

5 Multiclausal Utterances

In addition to the ability to partially specify information structure (a property shared with some previous approaches, including Kuhn (1996) and Song and Bender (2011)), the current proposal has the benefit of sufficient flexibility to handle multiclausal utterances. Specifically, the difference between our current proposal and our previous one is in the representation of the constraints: Where Song and Bender (2011) used features on semantic variables, here we introduce binary relations on ICONS in order to handle information structure in multiclausal sentences within the MRS representation.

(10)–(11) show how the move to binary relations helps represent cases where an individual has different information structural relations to the matrix and subordinate clauses. The answer in (10), which assigns the main stress (i.e. A-accent) on a constituent inside a relative clause, can be a proper answer to only Q1. Q2 is not a contextually appropriate question because it would require focus on the whole subject NP, and a non-head daughter (i.e. modifier) cannot project focus to its head daughter (i.e. modifiee) (Chung et al., 2003). In other words, [$_f$ *The dog that* KIM *saw*] is not a possible focus projection result because the head noun *dog* without an accent cannot inherit focus from KIM in the relative clause.[13] For the same reason, the answer sounds infelicitous in the *all-focus* context set up by Q3 as well. These facts suggest the range of focus projection possibilities shown in (11a). The encoding of these possibilities in our underspecified representation, along with further information structural information, is shown in (11b). The key property of (11b) is that one element *dog* is related via different elements of ICONS to two verbs; one is *barked* in the matrix clause, and the other is *saw* in the relative clause.[14] On the one hand, *dog* has the *non-focus* relation (i.e. either *topic* or *bg*) with the main verb *barked*, because it cannot inherit focus from the A-accent in the relative clause.[15] On the other hand, since there is no specific clue to identify the relation between *dog* and *saw*, *dog* is specified as just *info-str* in relation to *saw*. In (11b) there are three additional relations as well: On the one hand, KIM, which bears the A-accent (i.e. is overtly marked), has the *focus* relation with *saw* in the relative clause. On the other hand, *saw* and *barked* lack specific marking and so are left underspecified.

[13] If *dog* also bears the A-accent, it can get focus (i.e. multiple foci: *The* DOG *that* KIM *saw barked.*), but it cannot be focused through focus projection from the adjunct (Chung et al., 2003).

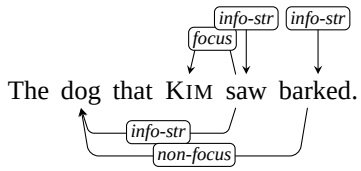
[14] The ICONS relationship between *dog* and *saw* is mediated by the coindexation of *dog* and the gap in the relative clause.

[15] Heycock (1994) and Chung et al. (2003) claim whether the focus on subjects can be projected to the whole sentence or not depends on an aspectual property of the predicates (i.e. individual-level *vs*. stage-level). Exploring naturally occurring texts, however, presents quite a number of examples which the distinction between individual-level and stage-level cannot be straightforwardly applied to. Thus, it would be more feasible to leave formally unmarked constituents (e.g. *barked* in (7b)) informatively underspecified.

(10) Q1: Which dog barked?
Q2: #What barked?
Q3: #What happened?
A: The dog that KIM saw barked.

(11) a. The dog that [$_f$ [$_f$ KIM] saw] barked.

b.



c.

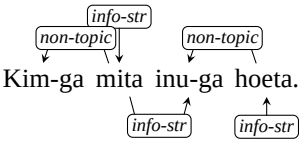
$$\begin{bmatrix} \textit{mrs} \\ \text{LTOP} \quad \boxed{h1} \\ \text{INDEX} \quad \boxed{e2} \\ \text{RELS} \quad \left\langle \begin{bmatrix} \textit{\_the\_q\_rel} \\ \text{LBL} \quad \boxed{h3} \\ \text{ARG0} \quad \boxed{x5} \\ \text{RSTR} \quad \boxed{h6} \\ \text{BODY} \quad \boxed{h4} \end{bmatrix}, \begin{bmatrix} \textit{\_dog\_n\_rel} \\ \text{LBL} \quad \boxed{h7} \\ \text{ARG0} \quad \boxed{x5} \end{bmatrix}, \begin{bmatrix} \textit{proper\_q\_rel} \\ \text{LBL} \quad \boxed{h8} \\ \text{ARG0} \quad \boxed{x10} \\ \text{RSTR} \quad \boxed{h9} \\ \text{BODY} \quad \boxed{h11} \end{bmatrix}, \right. \\ \left. \begin{bmatrix} \textit{named\_rel} \\ \text{LBL} \quad \boxed{h12} \\ \text{ARG0} \quad \boxed{x10} \\ \text{CARG} \quad \textit{kim} \end{bmatrix}, \begin{bmatrix} \textit{\_see\_v\_rel} \\ \text{LBL} \quad \boxed{h7} \\ \text{ARG0} \quad \boxed{e13} \\ \text{ARG1} \quad \boxed{x10} \\ \text{ARG2} \quad \boxed{x5} \end{bmatrix}, \begin{bmatrix} \textit{\_bark\_v\_rel} \\ \text{LBL} \quad \boxed{h14} \\ \text{ARG0} \quad \boxed{e2} \\ \text{ARG1} \quad \boxed{x5} \end{bmatrix} \right\rangle \\ \text{HCONS} \quad \left\langle \begin{bmatrix} \textit{qeq} \\ \text{HARG} \quad \boxed{h6} \\ \text{LARG} \quad \boxed{h7} \end{bmatrix}, \begin{bmatrix} \textit{qeq} \\ \text{HARG} \quad \boxed{h9} \\ \text{LARG} \quad \boxed{h12} \end{bmatrix} \right\rangle \\ \text{ICONS} \quad \left\langle \begin{bmatrix} \textit{focus} \\ \text{CLAUSE} \quad \boxed{e13} \\ \text{TARGET} \quad \boxed{x10} \end{bmatrix}, \begin{bmatrix} \textit{info-str} \\ \text{CLAUSE} \quad \boxed{e13} \\ \text{TARGET} \quad \boxed{e13} \end{bmatrix}, \begin{bmatrix} \textit{info-str} \\ \text{CLAUSE} \quad \boxed{e13} \\ \text{TARGET} \quad \boxed{x5} \end{bmatrix}, \right. \\ \left. \begin{bmatrix} \textit{non-focus} \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x5} \end{bmatrix}, \begin{bmatrix} \textit{info-str} \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{e2} \end{bmatrix} \right\rangle \end{bmatrix}$$

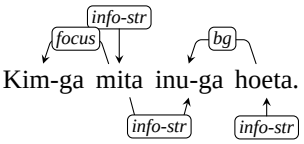
6 A Sample Translation

This section briefly illustrates how our representations are used in machine translation. The LOGON MT infrastructure (Oepen et al., 2007) handles translation in three steps: first, a sentence from the source language is parsed using the source language grammar, resulting in an MRS representation. Then that MRS is used as input to the transfer process where it is modified by transfer rules into an MRS interpretable by the target language grammar. Finally, the target language MRS is given to the generator, along with the target language grammar, and the generator finds realizations (surface strings) which the grammar licenses as compatible with the MRS.

(11c) above shows the full, underspecified MRS representation of (11a). The simpler graph-based view is in (11b). One potential translation of this sentence into Japanese is (12a). If we parse (12a) with the Japanese grammar, the resulting ICONS constraints are as in (12b). This is compatible with (11b,c). Thus, if we were to put (11c) through the transfer component, with an appropriate transfer grammar to map the English predicates to Japanese ones (leaving the ICONS intact), the resulting MRS could be used as input by the generator with the Japanese grammar to generate (12a).

(12) a. Kim-ga mita inu-ga hoeta
Kim-NOM saw dog-NOM barked [jpn]

b.


c.


In the process of generation, information from the input MRS is unified with constraints provided by the grammar. Thus the actual ICONS value associated with (12a) as translation output from (11a) will be the more specific representation shown in (12c). The *focus* relation between *Kim* and *mita* 'saw', which is a more specific type of *non-topic*, is taken from (11c). *Non-focus* between *dog* and *barked* in (11c) and *non-topic* between *inu* 'dog' and *hoeta* 'barked' are consistent with each other, and unified as *bg*. The others are the same as those in (12b).

7 Implementation

We have actually implemented the analyses discussed so far with ACE (`http://sweaglesw.org/linguistics/ace`).[16]

As the first step, we created toy grammars for English and Japanese using the LINGO Grammar Matrix customization system. Next we added the type hierarchy of ICONS and the related constraints into each grammar, which include (4), (5), and language-specific rules to mark information structure (i.e. A/B-accents in English, and lexical markers in Japanese). Using ACE, we conducted a small experiment to check out whether our grammars provide the translations as expected. For example, the English words *dog* and *barks* can bear the different ICONS values shown in (13), depending on their associated accents. We represent these accents with the hypothetical suffixes '-a' and '-b'. The '-b' suffix cannot be attached to the verb *barks* in our toy grammar because verbs presumably cannot be marked via B-accent for the information structural role of *topic* in English.

[16] ACE, using DELPH-IN grammars (such as the ERG (Flickinger, 2000) or grammars output by the Grammar Matrix customization system), parses sentences of natural languages, and generates sentences based on the MRS representation that the parser creates. It is the first DELPH-IN processor to specifically handle ICONS as part of the MRS.

(13) dog dog: info-str [ICONS: < e2 info-str x4 >]
dog-a: focus [ICONS: < e2 focus x4 >]
dog-b: topic [ICONS: < e2 topic x4 >]
bark barks: info-str [ICONS: < e2 info-str e2 >]
barks-a: focus [ICONS: < e2 focus e2 >]

Thus, *The dog barks* without any information structural marking logically can be interpreted as six types of sentences (3×2). However, if we apply ICONS to generation, we can filter out sentences which are not informatively equivalent to the input sentence. For example, if the input sentences are *The* DOG *barks* and *The **dog** barks* in which the subject bears the A/B-accents respectively, they can be monolingually paraphrased as (14). That is, we can get rid of two infelicitous sentences from each set of sentences.

(14) a. The dog-**a** barks [ICONS: < e2 **focus** x4, e2 info-str e2 >]
(i) The dog barks
(ii) The dog-a barks
(iii) The dog barks-a
(iv) The dog-a barks-a
(v) ~~The dog-b barks~~
(vi) ~~The dog-b barks-a~~
b. The dog-**b** barks [ICONS: < e2 **topic** x4, e2 info-str e2 >]
(i) The dog barks
(ii) ~~The dog-a barks~~
(iii) The dog barks-a
(iv) ~~The dog-a barks-a~~
(v) The dog-b barks
(vi) The dog-b barks-a

The same goes for Japanese in which lexical markers play a role to signal information structure. There are at least three Japanese translations (i.e. case-marking, topic-marking, and null-marking) corresponding to *The dog barks*, but case-marked NPs cannot be paraphrased into topic-marked NPs within our *info-str* hierarchy given in (3), and vice versa.

(15) a. inu **ga** hoeru [ICONS: < e2 **non-topic** x4, e2 info-str e2 >]
(i) inu ga hoeru
(ii) ~~inu wa hoeru~~
(iii) inu hoeru
b. inu **wa** hoeru [ICONS: < e2 **topic** x4, e2 info-str e2 >]
(i) ~~inu ga hoeru~~
(ii) inu wa hoeru
(iii) inu hoeru

Translating across languages is constrained in the same manner. An English sentence (16a) cannot be translated into (16a-ii), because the *focus* role that DOG involves is incompatible with the *topic* role that the topic maker *wa* assigns. On the other hand, a Japanese sentence (16b) cannot be translated into (16b-v) and (16b-vi), because *non-topic* that comes form the nominative marker *ga* is contradictory to *topic* that the B-accent signals in English.

(16) a. The dog-**a** barks [ICONS: < e2 **focus** x4, e2 info-str e2 >]
(i) inu ga hoeru
(ii) ~~inu wa hoeru~~
(iii) inu hoeru

b. inu **ga** hoeru [ICONS: < e2 **non-topic** x4, e2 info-str e2 >]
(i) The dog barks
(ii) The dog-a barks
(iii) The dog barks-a
(iv) The dog-a barks-a
(v) ~~The dog-b barks~~
(vi) ~~The dog-b barks-a~~

In our small experiment, we conducted four types of translation or paraphrasing (English-English, Japanese-Japanese, English-Japanese, and Japanese-English), and found that for our simple example sentences incorporating information structure into the translation process reduces the number of outputs by 22%.

8 Conclusion

This paper, in the context of multilingual MT, shows that information structure can be effectively represented within MRS via ICONS. ICONS takes as its value a list of *info-str* objects with CLAUSE and TARGET properties; the subtypes of *info-str* indicate which information structural role is played by the TARGET with respect to the CLAUSE.

Our future work includes two directions: Theoretically, it is important to understand how information structure works in various types of embedded clauses (e.g. clefts, control constructions) as well as what kinds of embedded constituents create their own information structural domains (e.g. relative clauses *vs*. progressive participles used as modifiers). Distributionally, we plan to exploit multilingual parallel texts to learn whether ICONS can be straightforwardly applied to other languages from a cross-linguistic viewpoint.

References

Bender, Emily M., Drellishak, Scott, Fokkens, Antske, Poulson, Laurie and Saleem, Safiyyah. 2010. Grammar Customization. *Research on Language & Computation* 8(1), 23–72.

Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Bildhauer, Felix and Cook, Philippa. 2010. German Multiple Fronting and Expected Topichood. In Stefan Müller (ed.), *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, pages 68–79, Stanford: CSLI Publications.

Choi, Hye-Won. 1999. *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI Publications.

Chung, Chan, Kim, Jong-Bok and Sells, Peter. 2003. On the Role of Argument Structure in Focus Projections. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 39, pages 386–404, Chicago Linguistic Society.

Copestake, Ann. 2002. Definitions of Typed Feature Structures. In Stephan Oepen, Dan Flickinger, Jun-ichi Tsujii and Hans Uszkoreit (eds.), *Collaborative Language Engineering*, pages 227–230, Stanford, CA: CSLI Publications.

Copestake, Ann. 2009. Slacker Semantics: Why Superficiality, Dependency and Avoidance of Commitment can be the Right Way to Go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 1–9, Athens, Greece: Association for Computational Linguistics.

Copestake, Ann., Flickinger, Dan., Pollard, Carl. and Sag, Ivan A. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation* 3(4), 281–332.

De Kuthy, Kordula. 2000. *Discontinuous NPs in German – A Case Study of the Interaction of Syntax, Semantics and Pragmatics*. CSLI publications.

Engdahl, E. and Vallduví, E. 1996. Information Packaging in HPSG. *Edinburgh Working Papers in Cognitive Science* 12, 1–32.

Flickinger, Dan. 2000. On Building a More Efficient Grammar by Exploiting Types. *Natural Language Engineering* 6(1), 15–28.

Gryllia, Styliani. 2009. *On the Nature of Preverbal Focus in Greek: a Theoretical and Experimental Approach*. Ph. D.thesis, Leiden University.

Hedberg, Nancy and Sosa, Juan M. 2007. The Prosody of Topic and Focus in Spontaneous English Dialogue. In Chungmin Lee, Matthew Gordon and Daniel Búring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, pages 101–120, Dordrecht: Kluwer Academic Publishers.

Heycock, Caroline. 1994. Focus Projection in Japanese. In *Proceedings North East Linguistic Society*, volume 24.

Huang, C.-T. James. 1984. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry* 15(4), 531–574.

Ishihara, Shinichiro. 2001. Stress, Focus, and Scrambling in Japanese. *MIT Working Papers in Linguistics* 39, 142–175.

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar.*. Cambridge, MA.: The MIT Press.

Kuhn, Jonas. 1996. An Underspecified HPSG Representation for Information Structure. In *Proceedings of the 16th conference on Computational Linguistics*, volume 2, pages 670–675, Association for Computational Linguistics.

Lambrecht, Knud. 1996. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge, UK: Cambridge University Press.

Neeleman, Ad and Titov, Elena. 2009. Focus, Contrast, and Stress in Russian. *Linguistic Inquiry* 40(3), 514–524.

Nguyen, Hoai Thu Ba. 2006. *Contrastive Topic in Vietnamese: with Reference to Korean*. Ph. D.thesis, Seoul National University.

Oepen, Stephan, Velldal, Erik, Lønning, Jan T., Meurer, Paul, Rosén, Victoria and Flickinger, Dan. 2007. Towards Hybrid Quality-Oriented Machine Translation – On linguistics and probabilities in MT. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde, Sweden.

Ouhalla, Jamal. 1999. Focus and Arabic Clefts. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 335–359, Amsterdam/Philadelphia: John Benjamins Publishing Co.

Paggio, Patrizia. 2009. The Information Structure of Danish Grammar Constructions. *Nordic Journal of Linguistics* 32(01), 137–164.

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL: The University of Chicago Press.

Siegel, Melanie. 1999. The Syntactic Processing of Particles in Japanese Spoken Language. In Jhing-Fa Wang and Chung-Hsien Wu (eds.), *Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation*, pages 313–320.

Song, Sanghoun and Bender, Emily M. 2011. Using Information Structure to Improve Transfer-based MT. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 348–368, Stanford: CSLI Publications.

Steedman, Mark. 2000. Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry* 31(4), 649–689.

Wee, Hae-Kyung. 2001. *Sentential Logic, Discourse and Pragmatics of Topic and Focus*. Ph. D.thesis, Indiana University.

Wilcock, Graham. 2005. Information Structure and Minimal Recursion Semantics. *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday* pages 268–277.