Using Information Structure to Improve Transfer-based MT

Sanghoun Song and Emily M. Bender

University of Washington

Proceedings of the HPSG 2011 Conference

Department of Linguistics, University of Washington

Stefan Müller (Editor)

2011

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper hypothesizes that transfer-based machine translation systems can be improved by encoding information structure in both the source and target grammars, and preserving information structure in the transfer stage. We explore how information structure can be represented within the HPSG/MRS formalism (Pollard and Sag, 1994; Copestake et al., 2005) and how it can help refine multilingual MT. Building upon that framework, we provide a sample translation between English and Japanese and check the feasibility of the proposals in small-scale translation systems built with the HPSG/MRS-based LOGON MT infrastructure (Oepen et al., 2007). Our experiment shows the information structure-based MT system that we propose in this paper reduces the number of translations 75.71% for Japanese and 80.23% for Korean. The dramatic reductions in the number of translations is expected to make a contribution to our HPSG/MRS-based MT in terms of latency as well as accuracy.

1 Introduction

In the context of MT, we find that allosentences – close paraphrases which share truth conditions (Lambrecht, 1996) – are not always felicitous as translations of the same inputs. For example, a simple English sentence (1a) can be translated into at least two Japanese allosentences such as (1b) (i.e. with the nominative marker *ga* or with the topic marker *wa*).

(1) a. I am Kim. (English)

b. watashi-ga/wa Kim desu.
I-NOM/TOP Kim COP [jpn]

However, the choice between the alternatives shown in (1) is conditioned by the given context; the NP marking hinges on whether or not *watashi* 'I' functions as the topic. If the sentence is an answer to a question like 'Who are you?', the topic marker *wa* is strongly preferred. In contrast, if the sentence is used in reply to a question like 'Who is Kim?', the answer with the topic marker *wa* sounds unnatural to Japanese native speakers.[1]

The difference in felicity conditions between allosentences is the subject of study of information structure. Thus, we hypothesize that information structure

[†]We thank Tim Baldwin, Dan Flickinger, Stephan Oepen, Francis Bond, Ann Copestake, Laurie Poulson, Antske Fokkens, Joshua Crowgey, Michael Wayne Goodman, Naoko Komoto, Jong-bok Kim, and Stefan Müller for comments and suggestions at various stages and to three anonymous reviewers for helpful feedback. 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.

[1]Japanese judgments reported in this paper were provided by Naoko Komoto.

can be used to improve machine translation. Information structure is hypothesized to be universal: All languages have some way to mark topics and foci, such as with pitch accent, word order, morphological marking or some combination of these (Gundel, 1999), though the marking is not necessarily unambiguous. The universality of information structure suggests that it should transfer well and that it in turn can help facilitate transfer when the syntactic structures in the source and target languages diverge.

The underlying hypothesis of this study is that translation is, in essence, the process of reshaping the means of conveying information, instead of simply changing the words or reordering phrases. Building upon this fundamental premise, this study sets up a working hypothesis: Transfer-based MT systems can be strongly supported by (i) encoding information structure in both the source and target grammars, and by (ii) preserving information structure in the transfer stage. That implies that information structure needs to be marked within the MRS representation in each step of the translation process: parsing, transfer, and generation.

In this paper, we explore (i) how information structure can be represented within the HPSG (Pollard and Sag, 1994) and Minimal Recursion Semantics (MRS; Copestake et al., 2005) formalisms and also (ii) how information structure can be used to improve our multilingual MT system. We also offer (iii) an experimental result to show the computational feasibility with a pair of small-scale MT systems built with the LOGON MT infrastructure (Oepen et al., 2007). This paper looks at the particular case of translating English passive sentences into Japanese and Korean. This case is of interest because active/passive pairs can yield relatively larger numbers of allosentences.

This paper is structured as follows: Section 2 provides a more concrete example which shows why it is necessary to look into information structure in the study of MT. Section 3 proposes a way to capture information structure in HPSG/MRS for the purpose of transfer-based MT. Section 4 covers how information structure is modeled in our source and target languages (English and Japanese/Korean, respectively) with the formalism given in Section 3. Section 5, next, shows how information structure can be used to refine translations with a sample translation, and measures the improvement that our system provides over a baseline system which does not refer to information structure in MT. Section 6 summarizes the paper and outlines plans for future work.

2 Basic Data

One type of example exhibiting structural divergence across languages in translation is active/passive pairs. In English, passives are used productively and constraints on passivization are relatively weak. In contrast, Japanese and Korean, which tend to downplay the role of passives, have stronger constraints on pas-

sivization.[2] Consider the Japanese sentences in (3), which are translations of the English sentences in (2). The active sentence (3a) is just fine, but the passive sentence (3b) sounds like a clumsy translation, as inanimate nouns tend not to appear in subject position of passive clauses in Japanese. That is, passives in one language cannot always be translated into passives in another. Though the syntactic encoding is different, the active sentence (3a) is one potentially legitimate translation of the English passive one (2b), while the passive one (3b) is not.

(2) a. Kim tore the book.
b. The book was torn by Kim. (English)

(3) a. Kim-ga sono hon-o yabut-ta.
Kim-NOM DET book-ACC tear-PST
'Kim tore the book.'
b. ?sono hon-ga Kim-ni yabu-rare-ta.
DET book-NOM Kim-DAT tear-PASS-PST
'The book was torn by Kim.' [jpn]

Moreover, even though transfer-based MT with semantic representations as the transfer level can translate the passive sentence (2b) into an active sentence in Japanese, there still remain two additional issues in translating English passives into Japanese. As presented in (1), case makers (e.g. *ga* for nominatives and *o* for accusatives) in Japanese are in complementary distribution with the topic marker *wa*. In addition, so-called scrambling (OSV order) is highly productive in Japanese (Ishihara, 2001); (4a) exhibits 'normal' major constituent order while (4b) illustrates scrambling, as the object *sono hon* 'the book' is followed by the subject 'Kim'. Hence, (3a) has at least eight allosentences ($2\times2\times2$) as given in (4).[3]

(4) a. Kim-ga/wa sono hon-o/wa yabut-ta.
Kim-NOM/TOP DET book-ACC/TOP tear-PST
b. sono hon-o/wa Kim-ga/wa yabut-ta.
DET book-ACC/TOP Kim-NOM/TOP tear-PST

[2] In fact, passive is not such a widespread phenomenon; Siewierska (2011) reports in WALS Online that languages without passives outnumber those with passives, showing a ratio of 211 to 162. This is consistent with the observation that the productivity of passivization differs in different languages, and underscores the need to be able to translate passives into actives and vice versa.

[3] An anonymous reviewer noted two facts regarding these allosentences. First, the so-called double *wa* construction, in which the topic marker *wa* attaches to both the subject and the object, occurs only rarely in Japanese. On the other hand, it is also true the double *wa* construction is not illegitimate in Japanese, though its productivity is rather low. We assume that the first *wa*-marked NP in a sentence is the topic of the sentence, and the second *wa*-marked NP conveys the meaning of contrastive-focus. Second, since Japanese allows so-called 'pro-drop', we can consider one more option. That is, *Kim* and *sono hon* 'the book' can be freely dropped, in appropriate discourse contexts. Moreover, since NP markers (e.g. *ga* and *wa*) are optional in Japanese, we have at least 32 allosentences in total. However, in this paper, as our aim is to verify whether or not information structure can improve performance of transfer-based MT with a small-scale experiment, we provisionally ignore these last two options.

What needs to be taken into consideration here is that these eight sentences are not felicitous in the same contexts, though they presumably share the same truth conditions. We propose to take sets of translation candidates like these (for more details, see §5) and refine them on the basis of information structure. In order to do so, we first explore how to represent information structure in MRS and then how to build those representations compositionally in HPSG grammars.

3 Information Structure in HPSG/MRS

Because assignment of information structure categories to referents can be constrained by both lexical marking and phrase-structural configurations, we analyze information structure in terms of three levels of structure: a semantic feature INFO-STR in the MRS (§3.1), a syntactic feature MKG encoding the lexical marking (§3.2), and a set of constraints on phrase structure rules relating the two (§3.3).

Our analysis builds on the following assumptions: First, while sentences always have at least one focus, they do not always have a topic (Gundel, 1999); further, constituents may be 'background' (i.e. neither topic nor focus) (Büring, 1999). Second, we treat 'contrast' as a cross-cutting information structure category, which contributes the entailment of an alternative set (Molnár, 2002). Lambrecht (1996) regards 'contrastiveness' as a merely cognitive concept, yet there are several cross-linguistic counterexamples to his claim; some languages employ specific markers or syntactic means to express contrastiveness. For example, Vietnamese uses a contrastive-topic marker *thì*, exemplified in (5) (Nguyen, 2006, p. 1). This marker is distinct from the regular topic marker (i.e. our *aboutness-topic*). The contrast function is shown by the alternative set evoked in (5), while the distinctiveness from focus is shown by the fact that *thì*-marked NPs cannot be used to answer *wh*-questions (*Ibid.*).

(5) Nam thì di Hanoi
Nam CT go Hanoi
'Nam goes to Hanoi(, but nobody else).' [vie]

We can also find syntactic marking of contrast in several languages. In Standard Arabic, for instance, contrastively focused items are normally preposed to the initial position of the sentence, while non-contrastively focused items which convey 'new information' (i.e. *semantic-focus* in this paper) are in-situ with a specific pitch accent, as exemplified in (6a-b) respectively (Ouhalla, 1999, p. 337).

(6) a. RIWAAYAT-AN ʔallat-at Zaynab-u
novel-ACC wrote-she Zaynab-NOM
It was a NOVEL that Zaynab wrote.

b. ʔallat-at Zaynab-u RIWAAYAT-an
wrote-she Zaynab-NOM novel-ACC
Zaynab wrote a NOVEL. [arb]

Similarly, in Portuguese, contrastive focus precedes the verb, while non-contrastive focus follows the verb (Ambar, 1999). In Russian, contrastive focus is preposed, while non-contrastive focus shows up clause-finally (Neeleman and Titov, 2009). In addition to these distributional facts, there is also evidence that contrast behaves differently from non-contrastive focus (or topic) in the semantics. On the one hand, regarding the difference between contrastive focus and non-contrastive focus, Gundel (1999) argues the former cannot have an effect on the truth conditions, whereas the latter is truth-conditionally relevant. On the other hand, Nakanishi (2007), who compares contrastive topic with non-contrastive topic (i.e. *aboutness-topic* in this paper) in Japanese, claims they can have a different scopal interpretation when they co-occur with negation.

Our third assumption is that semantically empty categories (e.g. complementizers, expletives) and syncategorematic items (e.g. relative pronouns) are informatively empty as well (i.e. assigned no information structure category, though they may be required by constructions which serve to mark information structure, such as the cleft construction in English). For example, in (7a), the expletive *it* and the copula *is* are semantically empty and the relative pronoun *that* is syncategorematic; thus, they are informatively vacuous. Likewise, since the preposition *by* in English passive sentences is assumed to be semantically void, it cannot take part in information structure, as shown in (7b).

(7) a. ~~It~~ ~~is~~ the book ~~that~~ was torn by Kim.

b. The book ~~was~~ torn ~~by~~ Kim.

Finally, we assume the canonical position of topics is sentence-initial in our sample of languages (English, Japanese, and Korean), though this generalization does not hold for all languages (Erteschik-Shir, 2007).

3.1 MRS: *info-str*

Although information structure is strictly speaking pragmatic rather than semantic, we represent it in our MRS semantic representations. Our motivation for doing so is primarily practical: The MT infrastructure we are using (Oepen et al., 2007) does MRS-based transfer. Thus, (contra Engdahl and Vallduví (1996), Bildhauer (2007), and Paggio (2009)), we encode information structure in the semantics (MRS) rather than in a CONTEXT attribute. Like Paggio, we associate information structure with semantic indices; however, while Paggio has information structure-related lists in the CONTEXT structure taking indices as their elements, we represent information structure with a feature on indices directly in the MRS. This feature (INFO-STR) draws its values from the hierarchy in Figure 1.[4]

[4] In associating information structure with indices alone, rather than as a relationship between an index and a particular clause, we are not fully accounting for how information structure works in multi-clausal sentences. We leave a more complete representation of information structure which encodes such relationships to future work.

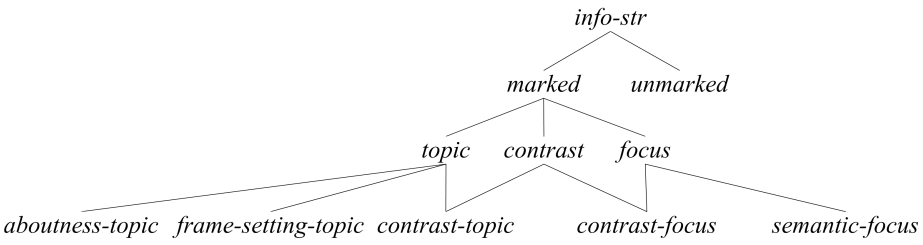

Figure 1: Type Hierarchy of *info-str*

Aboutness-topic refers to regular topics lacking a contrastive interpretation. *Frame-setting-topic* refers to adverbial expressions which present dimension of evaluation, such as 'as for' constructions in English or temporal/spatial adverbials which appear sentence-initially (Krifka, 2008). *Contrast-topic* and *contrast-focus* convey a contrastive interpretation, while *semantic-focus*, which does not introduce an alternative set, does not.

3.2 Markedness: *mkg*

The lexical marking itself is recorded via a syntactic feature MKG, inside of CAT. MKG has two subfeatures, TP and FC, which can be constrained independently.[5] The value of MKG is always a subtype of *mkg*, drawn from the hierarchy in Figure 2 (*Tp* is constrained to be [TP +], *non-tp* [TP −], *fc* [FC +], and *non-fc* [FC −]).

(8)
$$\begin{bmatrix} \text{MKG} & \begin{bmatrix} \text{TP} & \textit{bool} \\ \text{FC} & \textit{bool} \end{bmatrix} \end{bmatrix}$$

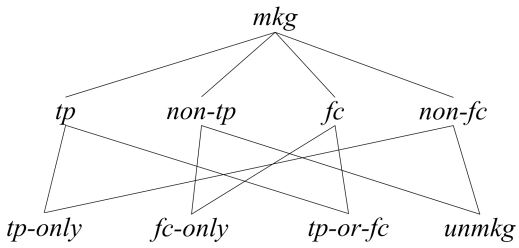

Figure 2: Type Hierarchy of *mkg*

The MKG value reflects the morphological marking but not necessarily the actual INFO-STR value because in some languages syntactic constructions assign the

[5]We believe that *mkg* could in principle be used in modeling focus projection, in the sense that foci can be classified into narrow focus and wide focus. Pursuing these ideas is left for future work.

INFO-STR, taking into account both the MKG value of the daughters and construction-specific constraints on their order. For instance, the topic markers *wa* in Japanese and *(n)un* in Korean can involve a focus reading if the topic-marked NP is scrambled as shown in (4b), which will be explained in detail in §4.2.

3.3 Sentential Forms: *sform*

Building on previous literature (Lambrecht, 1996; Engdahl and Vallduví, 1996; Paggio, 2009), we propose the classification of phrase types in Figure 3. *Topicality* is mainly concerned with how the topic is realized in a sentence. In *topic-comment* constructions (e.g. 'as for' constructions such as (9)), topics are followed by other constituents.[6]

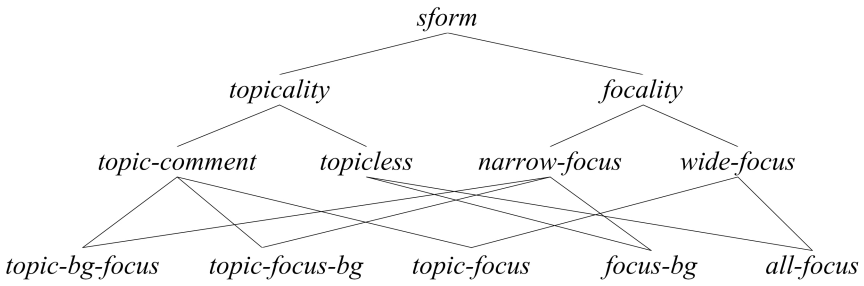

Figure 3: Type Hierarchy of *sform*

(9) As for **the book**, KIM tore it.

As noted, not all sentences have topics. We provide for this with the type *topicless* (e.g. cleft sentences in English such as (7a)). *Focality* is divided into *narrow-focus* and *wide-focus*. The distinction between them, however, is not necessarily equivalent to argument focus *vs*. predicate focus (Lambrecht, 1996; Erteschik-Shir, 2007), because verbs can bear *narrow-focus*.

Several of these sentence types are illustrated in English allosentences in (10)–(11), where we have added annotations disambiguating the information structure: SMALL CAPS for an A-accented phrase (H*), **boldface** for a B-accented one (L+H*), and $[_f\ ]$ for focus projection (Bolinger, 1961; Jackendoff, 1972).

(10) a. **The book** was torn by $[_f$ KIM].
b. **The book** $[_f$ was torn] by Kim.
c. **The book** $[_f$ was torn by Kim].
d. $[_f$ THE BOOK] was torn by Kim.
e. $[_f$ The book was torn by Kim].

[6]The conventions used in (9) are described above (10).

(11) a. **The book** [$_f$ was torn].
b. [$_f$ THE BOOK] was torn.
c. [$_f$ The book was torn].

In (10a), the subject 'the book' has a B-accent and the agent 'Kim' that follows the verb bears an A-accent (i.e. argument focus), which correspond to *topic* and *focus* of the sentence, respectively. As the remaining part 'was torn by', which is neither of them, corresponds to *bg*, we find (10a) is encoded as *topic-bg-focus* in the order named, which is the most unmarked *sform* in English (Lambrecht, 1996).[7] (10b-c), with predicate foci, are *topic-focus-bg* and *topic-focus*, respectively. The focus 'was torn' (i.e. narrow focus on the verb) is followed by the background 'by Kim' in (10b), unlike (10a). (10c) with *wide-focus* does not include any background. (10d-e) are *topicless*; *focus-bg* and *all-focus*, respectively. The cleft sentence (7a) is virtually the same as (10d) in terms of *sform*, because the expletive 'it', the copula, and the relative pronouns in clefts are informatively empty. That is, all cleft constructions in English are instances of *focus-bg*. *All-focus* (a.k.a. sentence focus in which the entire sentence is asserted) is typically an answer to the question like 'What happened?' (Lambrecht, 1996). On the other hand, the agent often disappears in passive sentences, as shown in (11). Since *topic-bg-focus* and *topic-focus-bg* that require three components are ruled out from (11) consisting of only the subject and the verb, there are three readings; *topic-focus* for (11a), *focus-bg* for (11b), and *all-focus* for (11c).

If INFO-STR is lexically or prosodically determined as in (10a), SFORM can be easily detected as well. For example, the 'as for' construction in English, such as (9), belongs to *topic-comment* because the (near) lexical expression 'as for' which has the *tp-only* (i.e. [TP +, FC –]) marks (*contrastive*)-*topic*, and the NP precedes the comment; (9) is encoded as *topic-focus-bg*. However, since the Japanese marker *wa* itself is informatively ambiguous, the syntactic configuration is required to determine SFORM as well as INFO-STR of each sentence in (4), as discussed in the next section.

4 Information Structure in English and Japanese/Korean

4.1 English

In English, information structure is normally constrained by pitch accents (Bolinger, 1961; Jackendoff, 1972)[8]; thus, English uses the A-accent (H*) to prosodically

[7] In contrast, in head-final languages (e.g. Japanese and Korean) in which the most unmarked focus position is immediately preverbal *topic-focus-bg* is the most unmarked *sform* (Ishihara, 2001). This implies that the most unmarked sentential forms differ in different languages, being largely dependent upon the default word order (Lambrecht, 1996; Erteschik-Shir, 2007): First, subjects normally are the most unmarked topics in most languages. That means subjects mostly function as the topic of the sentence unless there is a special cue to identify topic. Second, it is cross-linguistically common that an object is a case of unmarked argument focus.

[8] We are not considering the pitch accents directly in this study.

mark foci and the B-accent (L+H*) to prosodically mark topics, as presented in (12). As for contrast in English, its prosodic marking is partially similar to both A/B-accent (Hedberg and Sosa, 2007). As a result, both accents can be interpreted as contrast, in an appropriate context. Therefore, we assign the INFO-STR values *topic* and *focus*, which are compatible with the more specific *contrast-topic* and *contrast-focus* as well as *aboutness-topic* and *semantic-focus*.

(12)
$$\textit{fp-lex-rule} \rightarrow \begin{bmatrix} \text{PROSODY A-}\textit{accent} \\ \text{INFO-STR } \textit{focus} \end{bmatrix} \qquad \textit{tp-lex-rule} \rightarrow \begin{bmatrix} \text{PROSODY B-}\textit{accent} \\ \text{INFO-STR } \textit{topic} \end{bmatrix}$$

In the context of our text-based MT, this property might be problematic, because written English does not explicitly mark prosody, removing this cue to information structure. However, information structure categories presumably could be added to an English input sentence as a preprocessing step, either on the basis of prosodic analysis in a speech-based system or on the basis of a classifier which takes extra- as well as intra-sentential context into account. For present purposes, we represent these patterns with typeface variations in this paper. In the evaluation process of this study, we tentatively made use of hypothetical suffixes '-TP', '-FP', which represent B-accent for topics, and A-accent for foci respectively. For instance, (10a) is entered into our system as 'The book-TP was torn by Kim-FP'.[9]

4.2 Japanese/Korean

Japanese and Korean employ topic markers (*wa* and (*n*)*un*, respectively) which actively participate in encoding information structure. The topic markers in Japanese and Korean can also be used to denote contrastiveness. For example, as exemplified in (13), the sentence with the topic marker *wa* can sometimes be a felicitous answer to a given question.

(13) Q: Who came?

A: Kim-ga/wa ki-ta.
Kim-NOM/TOP come-PAST
'Kim came.' [jpn]

Kim-ga/wa in (13) directly correspond to the wh-word in the given question,[10] which means 'Kim' has to be interpreted as the focus of the sentence though the topic marker *wa* is attached to it. This implies the lexical marking in Japanese

[9]English also uses lexico-syntactic patterns to mark information structure, notably clefts, English focus movement, and *as for*. As these are much less pervasive than prosodic marking of information structure in English (and morphosyntactic marking in Japanese and Korean), we leave the integration of these into our English grammar fragment for future work.

[10]Many previous studies employ wh-questions as diagnostics to identify focus (e.g. Partee, 1991; Lambrecht, 1996; Gundel, 1999).

does not necessarily directly constrain the information structure in the way that prosodic marking in English does. *Kim-ga*/*wa* in (13), however, do not have the same meaning as each other (i.e. *semantic-focus vs*. *contrast-focus*). In an actual sense, if the topic marker *wa* is made use of, the answer conveys the meaning like 'Kim surely came, but whether anybody else came or not lacks confirmation.' (14) shows the difference between them more clearly.

(14) Kim-ga/#wa ki-ta-si, Lee-mo ki-ta.
Kim-NOM/TOP come-PAST-and, Lee-also come-PAST.
'Kim came and Lee also came.' [jpn]

Contrast never shows up out of the blue, because it has to involve an exclusive selection from alternatives (i.e. an available contrast set in the given context). Thus, if 'Kim' is exclusively chosen with the topic marker *wa*, (14) in which the alternative 'Lee' co-occurs sounds awkward. In sum, *wa*-marked NPs can be interpreted as *contrast-focus*.

The lexical markers alone do not fully identify the information structure in Japanese and Korean. Further information comes from word order, and in particular the phenomenon of scrambling (e.g. (4b)) (Choi, 1999; Ishihara, 2001). Whereas scrambling in Japanese/Korean has often been considered as a syntactically optional, semantically void operation, Ishihara argues it is an operation that offers potential focus sets which are not available with different word orders. Assuming Reinhart (1995)'s Focus Rule[11], Ishihara claims that there is a set of constituents that can serve as a focus domain as exemplified in (15) taken from Ishihara (2001, p. 157). (15a) in which the object *hon-o* 'book-ACC' bears the main stress of the given sentence has the focus set as (15c), which means any syntactic constituent containing the stressed word (i.e. OBJ as an argument focus, VP as a predicate focus, and IP as a sentence focus) can be the focus of the sentence.

(15) a. Taro-ga hón-o kat-ta.
Kim-NOM book-ACC buy-PST
'Taro bought a book.'

b. $[_{IP}$ SUBJ $[_{VP}$ $[_{DP}$ OBJ] V]]

c. Focus Set = {OBJ, VP, IP}

However, if the sentence is scrambled as (16b) taken from Ishihara (2001, p. 159), the focus set is also computed differently; VP1 in (16b) cannot function as the focus of the sentence, because it does not include the stressed element.[12]

[11] The focus of IP is a(ny) constituent containing the main stress of IP, as determined by the stress-rule.

[12] According to Cinque (1993), the main stress in head-final languages (e.g. Japanese, Korean) has a strong tendency to fall on the preverbal phrase. For instance, the object *hon* 'book' is most likely to have the main stress in (16a), while *kyoo* 'today' bears it in (16b).

(16) a. [$_{IP}$ Taro-ga [$_{VP2}$ kyoo [$_{VP1}$ [$_{DP}$ hón-o] kat-ta]]]
Taro-NOM today book-ACC buy-PST
Focus Set = {OBJ, VP1, VP2, IP}

b. [$_{IP2}$ hon-o [$_{IP1}$ Taro-ga [$_{VP2}$ [$_{ADV}$ kyóo] [$_{VP1}$ kat-ta]]]]
book-ACC Taro-NOM today buy-PST
Focus Set = {ADV, VP2, IP1, IP2}

In a similar vein, Choi differentiates contrasts from non-contrastive foci and topics in Korean. First, contrasts can freely scramble, while non-contrastive foci (a.k.a. *semantic-focus* (Gundel, 1999)) cannot. Second, when (*n*)*un* attaches to the in situ (i.e. non-scrambled) subject, the subject can be either *aboutness-topic* or *contrast-topic*. On the other hand, when (*n*)*un* attaches in situ non-subjects (e.g. objects), such constituents have only the contrastive reading.

We note the following generalizations which appear to hold for both Japanese and Korean: First, as discussed above, the markers *wa* and (*n*)*un* do not directly constrain information structure, but rather interact with word order phenomena to do so. Second, constituents marked with *wa* or (*n*)*un* are however marked as not 'background' (i.e. topic or focus, contrastive or otherwise). Third, *wa* or (*n*)*un* cannot appear in *all-focus* constructions that allow only *semantic-focus* lacking contrastive meanings, as exemplified in (17).

(17) Q: What happened?
A: Kim-ga/#wa sono hon-o/#wa yabut-ta.
Kim-NOM/TOP DET book-ACC/TOP tear-PST

Finally, we note the three possible interpretations of a *wa*- or *(n)un*-marked NP, depending on its syntactic function and position, shown in Table 1 adapted from Choi (1999). Although (4) illustrates the range of possible translations in Japanese corresponding to the English passive sentence (2b), they have different information structure in accordance with Table 1, as given in (18).

Table 1: Information Structure of Topic-marked NP

	in-situ	**scrambling**
subject	*topic*	*contrast-focus*
non-subject	*contrast-focus*	*contrast-topic*

(18) a. Kim-wa sono hon-o yabut-ta.
Kim-TOP DET book-ACC tear-PST
(*topic*)

b. sono hon-o Kim-wa yabut-ta.
DET book-ACC Kim-TOP tear-PST
(*contrast-focus*)

c. Kim-ga sono hon-wa yabut-ta.
Kim-NOM DET book-TOP tear-PST
(*contrast-focus*)

d. sono hon-wa Kim-ga yabut-ta.
DET book-TOP Kim-NOM tear-PST
(*contrast-topic*)

In short, the challenge in Japanese and Korean is to map from the morphological marking in combination with phrase structure patterns to the specific INFO-STR, including *contrast-topic* and *contrast-focus* which are the only possible interpretations of topic-marked NPs in certain positions. To handle this, we first use MKG to associate partial information with the nominative and topic markers:

(19)
$$\begin{array}{ll} \textit{nom-marker} \rightarrow & \textit{topic-marker} \rightarrow \\ \begin{bmatrix} \text{ORTH} \left\langle \textit{ga} \right\rangle \\ \text{MKG } \textit{unmkg} \\ \text{CASE } \textit{nom} \end{bmatrix} & \begin{bmatrix} \text{ORTH} \left\langle \textit{wa} \right\rangle \\ \text{MKG } \textit{tp} \\ \text{CASE } \textit{case} \end{bmatrix} \end{array}$$

The value of MKG is mapped to values of INFO-STR via the constraints on the various *sform* types. *Topic-comment* requires *tp* of non-head-daughter such that only NPs with topic markers can participate in *topic-comment*. The construction itself is [MKG *tp*] so that constituents which have picked up a topic cannot serve as the head daughter of another *topic-comment* phrase.

(20)
$$\begin{bmatrix} \textit{topic-comment} & \\ \text{MKG} & \textit{tp} \\ \text{HD} \mid \text{MKG} & \textit{fc} \\ \text{NON-HD} \mid \text{MKG} & \textit{tp} \end{bmatrix}$$

In this way, INFO-STR in Japanese and Korean, unlike in English, is specified at the phrasal level (by grammatical rules, such as specialized subtypes of *subj-head* and *comp-head*). The phrasal rules are now classified into eight subrules, which inherit from two types of head-phrases (i.e. *subj-head-phrase* and *comp-head-phrase*) and optionally *topic-comment*. The type hierarchy is sketched in Figure 4, in which there are two factors that have an influence on branching nodes; topic-marking and scrambling.

On the one hand, four rules which the prefix *top* is attached to multiply inherit from *topic-comment* as well as either *subj-head-phrase* or *comp-head-phrase*. On the other hand, four rules that contains *scr* that stands for 'scrambled' deal with constructions in which the non-head-daughter is not in-situ. As presented in (21), INFO-STR in Japanese and Korean is specified in each rule. *Top-scr-subj-head* in

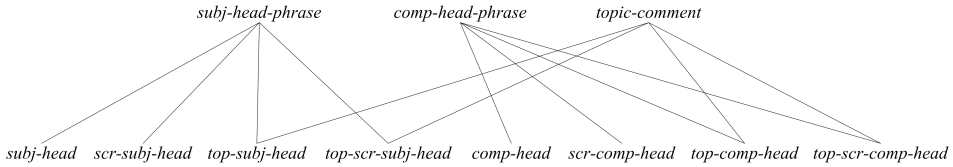

Figure 4: Type Hierarchy of Phrasal Rules

(21) specifies INFO-STR of the non-head-daughter (i.e. a subject) as *contrast-focus* in accordance with Table 1. The non-head-daughter in *top-scr-comp-head* (i.e. a non-subject), likewise, is specified as *contrast-topic*.

(21) $\begin{bmatrix} \textit{top-scr-subj-head} \\ \text{HD} \mid \text{VAL} \mid \text{COMPS} \left\langle [] \right\rangle \\ \text{NON-HD} \mid \text{INFO-STR} \ \textit{contrast-focus} \end{bmatrix} \begin{bmatrix} \textit{top-scr-comp-head} \\ \text{HD} \mid \text{VAL} \mid \text{COMPS} \langle\rangle \\ \text{NON-HD} \mid \text{INFO-STR} \ \textit{contrast-topic} \end{bmatrix}$

For example, Figure 5 shows the derivation tree of (22). The phrase structure rule building the node combining the subject and the verb for (22) (attaching *Kim-ga* 'Kim-NOM' to the rest of the sentence) is an instance of *scr-subj-head*, which combines via the *top-scr-comp-head* rule with the topic-marked object *sono hon-wa*.

(22) sono hon-wa Kim-ga yabut-ta.
DET book-TOP Kim-NOM tear-PST

NPs with nominative markers (e.g., *Kim-ga* in (22)) can't be interpreted as either topic or contrast (i.e., must be non-constrastive focus or background), because the non-head-daughter of *topic-comment* is incompatible with [TP –] as given in (20). On the other hand, *sono hon-wa* 'DET book-TOP' in (22) is a scrambled complement; it is licensed by *top-scr-comp-head* which inherits from both *comp-head-phrase* and *topic-comment*. Its INFO-STR is *contrast-topic* because of the constraint on the rule shown in (21). This models the fact that it is interpreted as both contrast and topic.

5 Translation

For our experiment, we made use of 24 input sentences in English; eight types of allosentences as shown in (10)–(11) for each of the three verbal types: 'tear', 'chase', and 'hit' as exemplified in Table 2 (i.e. 8×3). The first verbal type takes inanimate nouns as complements, and thus resists passivization in Japanese and Korean. The second one tends to be freely passivized. The third one does not have passive forms in Korean, whereas it can be passivized in Japanese. Table 2 compares the linguistic properties of source/target languages discussed so far, and gives three types of verbs in each language.

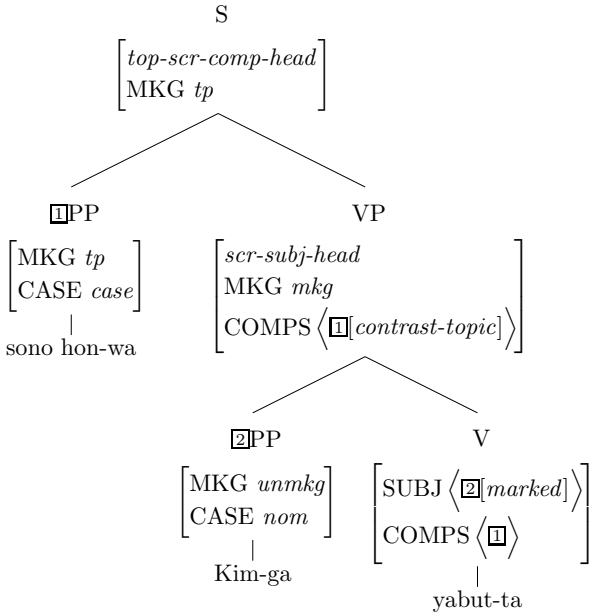

Figure 5: A Sample Derivation in Japanese

5.1 A Sample Translation

The most remarkable advantage of the model that we propose is that information structure-based system can significantly reduce the number of translations. Information structure in MT can function as a filter to reduce the number of candidate translations. To illustrate the process, we will step through the translation of (10a), which has at least eight potential translations in Japanese as given in (4), if we ignore information structure.

Parsing (English): The corresponding tree derivation is sketched out in Figure 6, in which 'the book' with the B-accent is straightforwardly specified as *topic*, and 'Kim' with the A-accent is specified as *focus*.

Transfer and Input/Output MRS: The transfer stage takes as its input the MRS in Figure 7, from the English parse tree, which specifies [INFO-STR *topic*] on the ARG0 of *_book_n_rel* (shared with the ARG0 of *exist_q_rel*), and [INFO-STR *focus*] on that of *named_rel* for 'Kim'. This information is preserved in the mapping to the target language MRS in Figure 8.[13]

Generation (Japanese): The Japanese grammar used in generation only generates structures which are compatible with the input MRS (Figure 8), including the constraints it places on INFO-STR. Because only *wa*-marked NPs can be topics in Japanese, *sono hon* 'the book' must be marked by *wa* in any realization of this

[13] In this study, we avoid the need for transfer rules by using pseudo-interlingual predicate names. This approach works at the very small scale we are experimenting at, but does not scale up. The LOGON system provides extensive support for developing transfer grammars.

Table 2: Source/Target Languages

	English	**Japanese**	**Korean**
focus	A-accent	case markers	
topic	B-accent	topic markers (*wa*, (*n*)*un*)	
contrast	A/B-accent		
passives	productive	less productive	
animacy	insensitive	sensitive	
verb1	'tear'	*yaburu-*	*ccic-*
verb2	'chase'	*ou-*	*ccoch-*
verb3	'hit'	*naguru-*	*ttayli-*

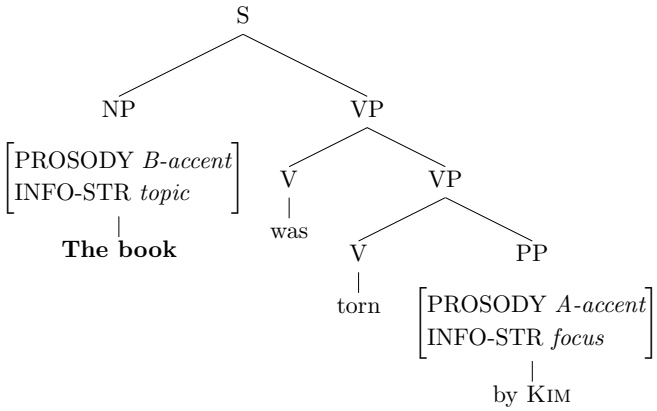

Figure 6: A Sample Derivation in English

MRS. Furthermore, since topics must be sentence-initial, only scrambled versions of the sentence are generated.

Using this constraint, now we can rule out infelicitous sentences. There are, as stated before, eight potential translations as given in (23): ~~strike~~ in (23) indicates the sentence is regarded as an inappropriate translation in the given context, and thus not generated by the grammar that takes information structure into account.

(23) a. ~~Kim-ga sono hon-o yabut-ta.~~
b. ~~Kim-ga sono hon-wa yabut-ta.~~
c. ~~Kim-wa sono hon-o yabut-ta.~~
d. ~~Kim-wa sono hon-wa yabut-ta.~~
e. ~~sono hon-o Kim-ga yabut-ta.~~
f. ~~sono hon-o Kim-wa yabut-ta.~~
g. sono hon-wa Kim-ga yabut-ta.
h. sono hon-wa Kim-wa yabut-ta.

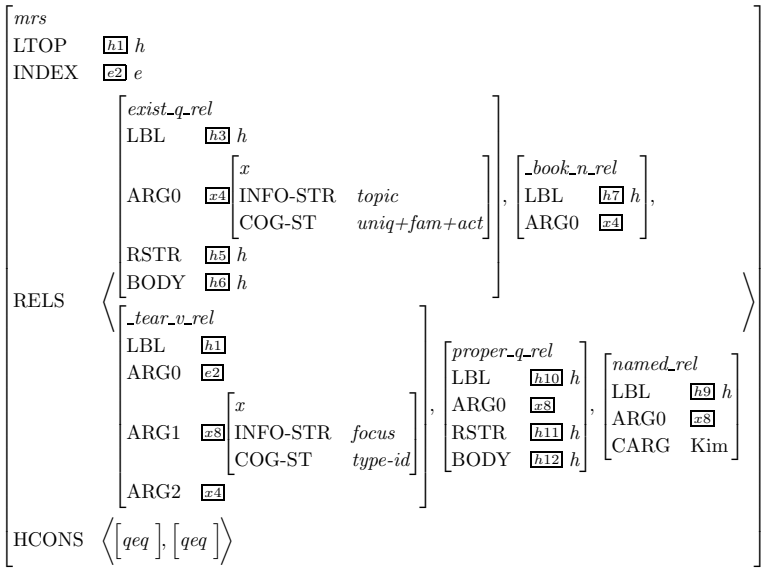

Figure 7: Input MRS (English)

First, since 'the book' is the topic and topics in Japanese must occur sentence-initially, (23a-d) are not generated. Second, (23e-f) in which *sono hon* is not topic-marked are not generated, because the *o*-marked NPs with [MKG *unmkg*] cannot be used as the non-head-daughter of *topic-comment*. Finally, when the underspecified value *focus* of 'Kim' in the MRS is passed to the Japanese grammar, the Japanese grammar provides two different outputs that are consistent with *semantic-focus* and *contrast-focus*, respectively. On the one hand, *ga*-marked *Kim* in (23g) is consistent with a context that calls for semantic focus but no contrast. On the other hand, *wa*-marked Kim in (23h) is interpreted as *contrast-focus* in accordance with Table 1. As a result, only the scrambled variants (23g-h) are generated as the felicitous translations directly corresponding to (10a). That is, we filter out 6 infelicitous translations out of 8 potential translations. For an example derivation, see Figure 5, which corresponds to (23g).

5.2 Evaluation: Translating Passives

To evaluate these proposals, we have implemented them in tdl (type description language), the high-level language interpreted by the LKB (Copestake, 2002). The first step is to construct small starter grammars for English, Japanese, and Korean, using the Grammar Matrix customization system (Bender et al., 2010). As a second step, other rules to produce allosentences (e.g. actives/passives) are added to each starter grammar. The third step is to implement information structure into each grammar, as given earlier. Finally, we create the mapping between internal and external features of indices (*semi.vpm*), in accordance with the LOGON MT

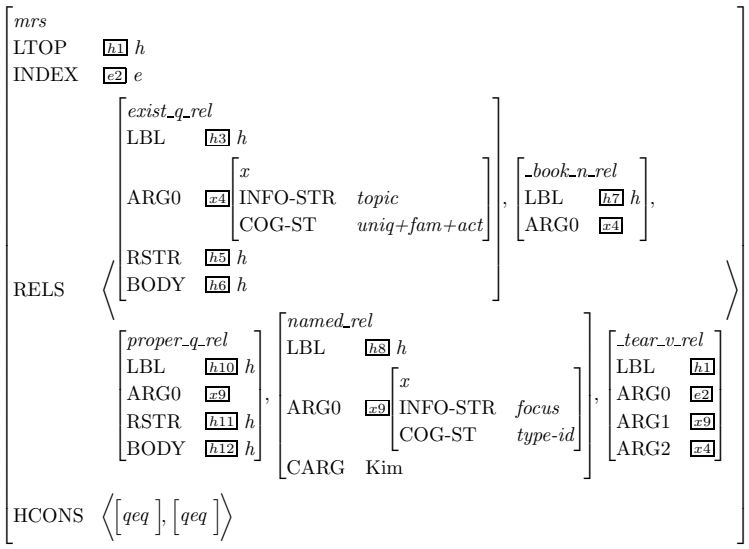

Figure 8: Output MRS (Japanese)

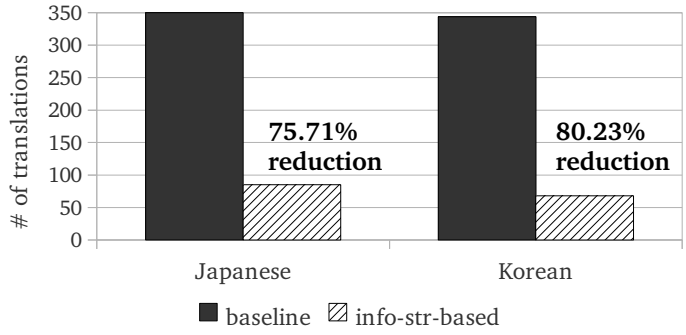

Figure 9: Evaluation

infrastructure (Oepen et al., 2007).

Our experiment shows our information structure-based system, compared to the baseline that lets all of potential translations through (without filtering for information structure), filters out 265 outputs in Japanese and 276 in Korean.[14] Consequently, as shown in Figure 9, we can reduce the number of outputs by 75.71% (from 350 to 85) for Japanese, and by 80.23% for Korean (from 344 to 68).

Thus, our information structure-based MT system has reduced the number of translations dramatically, which has two obvious effects on the performance of transfer-based MT: First, the processing burden of MT component which ranks the translations and select only suitable results can be greatly lightened, which

[14] We hand-verified the filtered Korean outputs and found that they were indeed less suitable.

should improve translation speed. Second, though it is still necessary to harness a re-ranking model for choosing translations, we can start from once-refined sets of translations, which should improve translation accuracy.

6 Conclusion

In this paper, we have made a proposal for how to represent information structure within the HPSG/MRS framework and have shown how it can be used to refine translations, especially focusing on translating English passives. The implications of this study are as follows: One the one hand, since the type hierarchies for information structure that this paper proposes are constructed almost language-independently, we are optimistic that they will apply to other language pairs as well. On the other hand, by enriching our semantic representations with information structure, we effectively move further up the MT pyramid (Vauquois, 1968), reducing the burden on the transfer component. Semantic-transfer based MT allows a system to handle a broad range of structural divergences. However, this also means that the search space of possible translations get larger. We expect information structure to be useful in navigating the array of possibilities provided by many different syntactic constructions and (thus types of syntactic divergence).

Our future work includes the following: First, we plan to evaluate our information structure-based system with various types of sentences, such as clefting, topicalized sentences, and topic-drop sentences. Second, other language pairs also need to be covered in order to check out the feasibility of this proposal. In particular, MT from Japanese/Korean to English has to be examined in the sense that Japanese/Korean employ more specific information structure than English in our proposal. Third, we plan to extend our analyses to handle information structure in multi-clausal sentences. Finally, we plan to build up an library of information structure analyses for the Grammar Matrix customization system (Bender et al., 2010), which contains and extends the main proposals of this paper.

References

Ambar, Manuela. 1999. Aspects of the Syntax of Focus in Portuguese. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 23–54, Amsterdam/Philadelphia: John Benjamins Publishing Co.

Bender, Emily M., Drellishak, Scott, Fokkens, Antske, Poulson, Laurie and Saleem, Safiyyah. 2010. Grammar Customization. *Research on Language & Computation* 8(1), 23–72, 10.1007/s11168-010-9070-1.

Bildhauer, Felix. 2007. *Representing Information Structure in an HPSG Grammar of Spanish*. Ph. D.thesis, Universität Bremen.

Bolinger, Dwight L. 1961. Contrastive Accent and Contrastive Stress. *Language* 37(1), 83–96.

Büring, Daniel. 1999. Topic. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 142–165, Cambridge: Cambridge University Press.

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

Cinque, Guglielmo. 1993. A Null Theory of Phrase and Compound Stress. *Linguistic Inquiry* 24(2), 239–297.

Copestake, Ann. 2002. *Implementing Typed Feature Structure Grammars*. Stanford, CA: CSLI Publications.

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

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

Erteschik-Shir, Nomi. 2007. *Information Structure: The Syntax-Discourse Interface*. USA: Oxford University Press.

Gundel, Jeanette K. 1999. On Different Kinds of Focus. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 293–305, Cambridge: Cambridge University Press.

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

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.

Krifka, Manfred. 2008. Basic Notions of Information Structure. *Acta Linguistica Hungarica* 55(3), 243–276.

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

Molnár, Valéria. 2002. Contrast – from a Contrastive Perspective. In H. Hasselgrd, S. Johansson, B. Behrens and C. Fabricius-Hansen (eds.), *Information Structure in a Cross-Linguistic Perspective*, pages 147–162, Amsterdam, Netherland: Rodopi.

Nakanishi, Kimiko. 2007. Prosody and Information Structure in Japanese: A Case Study of topic Marker wa. In Chungmin Lee, Matthew Gordon and Daniel Büring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, pages 177–193, Dordrecht: Kluwer Academic Publishers.

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.

Partee, Barbara H. 1991. Topic, Focus and Quantification. *Cornell Working Papers in Linguistics* 10, 159–187.

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

Reinhart, Tanya. 1995. Interface Strategies. *OTS Working Papers, Utrecht University* .

Siewierska, Anna. 2011. Passive Constructions. In Matthew S. Dryer and Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library.

Vauquois, Bernard. 1968. A Survery of Formal Grammars and Algorithms for Recognition and Transformation in Mechanical Translation. In *IFIP Congress (2)*, pages 1114–1122.