The Complexity of Scrambling in Japanese: A TAG Approach

Alexander Diez
Heinrich-Heine University Düsseldorf, Germany
Alexander.Diez@hhu.de

Abstract

In this article, I present Japanese local and long-distance scrambling and restrictions to this phenomenon. I will argue that Japanese scrambling is too complex to be adequately represented with TAG. Instead, I will use a variant of TAG, namely TL-MCTAG. Subsequently, I also will propose to regard other scrambling languages, such as German, or Russian, in complexity classes, which is basically driven by the derivational power of each TAG formalism. This classification, though remains peripheric.

1 Introduction

This paper focuses on scrambling in Japanese, a language well known for its relative free word order. As a strict SOV language, Japanese verbs are linearized at the right end of a VP, while the other constituents may precede in any order without changing the denotation of the VP. This sort of flexiblity is known as scrambling (Bailyn, 2002, 83). There are roughly two types of scrambling, namely LOCAL SCRAMBLING (LS) and LONG DISTANCE SCRAMBLING (LDS). LS permits free word order inside the domain of a governing verb. Besides of the canonical order (1a), any permutation of the constituents is possible. (1b-c) shows some of the possible permutations.

(1) a. Hanako-ga hon-o
Hanako-NOM book-ACC
otōto-ni ageta
little.brother-to give.PST
'Hanako gave the book to the little brother.'

b. Otōto-ni Hanako-ga hon-o ageta.

c. Hon-o otōto-ni Hanako-ga ageta.

In contrast, LDS is highly restricted. LDS extends constituent boundary beyond the domain of a governing verb, as shown in (2b) in contrast to the canonical order in (2a).

(2) a. Kaito-ga [kinō Tarō-ga
Kaito-NOM yesterday Taro-NOM
Ginza-de sushi-to-sashimi-o
Ginza-in sushi-and-sashimi-ACC
tabeta] to Hanako-ni itta.
eat.PST COMP Hanako-to say.PST
'Kaito said, that yesterday Tarō ate sushi and sashimi in Ginza.'

b. Sushi-to-sashimi-o$_i$ Kaito-ga [kinō Tarō-ga Ginza-de t$_i$ tabeta] to Hanako-ni itta.

TREE ADJOINING GRAMMAR (TAG) (Joshi and Schabes, 1997) is a formalism based on tree-rewriting, where the so-called elementary trees can be combined by two compositional operations, substitution for initial trees and adjunction for auxiliary trees, and generate larger trees, as shown in Fig. 1. Both initial trees of *Peter* and *the fridge* rewrite the non-terminal leaves (substitution) of the *repaired*-tree. The auxiliary *easily*-tree rewrites the inner VP node (adjunction) of the *repaired*-tree. Auxiliary trees have a distinguished non-terminal leaf, the footnode, marked by an asterisk (*). After adjunction, the subtree of the rewritten target node appears below the footnode. The result of substitution and adjunction is a unique, derived tree. If every elementary tree includes at least one lexical anchor, the grammar is called lexicalized, hence a LEXICALIZED TAG (LTAG). It is, furthermore, possible to enrich the nodes of an elementary tree with feature structures, which is then called FEATURE STRUCTURE BASED TAG (FTAG) (Vijay-Shanker and Joshi, 1988). Such a unification based system is as powerful as TAG, but enhances the 'descriptive' capacity. Still, the generative power of

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 1–7,
Hissar, Bulgaria, 7-9 September 2015.

LTAG is not enough for Japanese LDS. This task, as will be shown in Section 3.3, can be solved with TREE-LOCAL MULTI-COMPONENT TAG (TL-MCTAG) (Weir, 1988). TL-MCTAG consists of sets of elementary trees, which must adjoin or substitute into one and the same elementary tree, and is equal to LTAG in terms of generative power. In section 3.3 I will show that TL-MCTAG, in contrast to LTAG, has the desired power to adequately analyse Japanese scrambling.

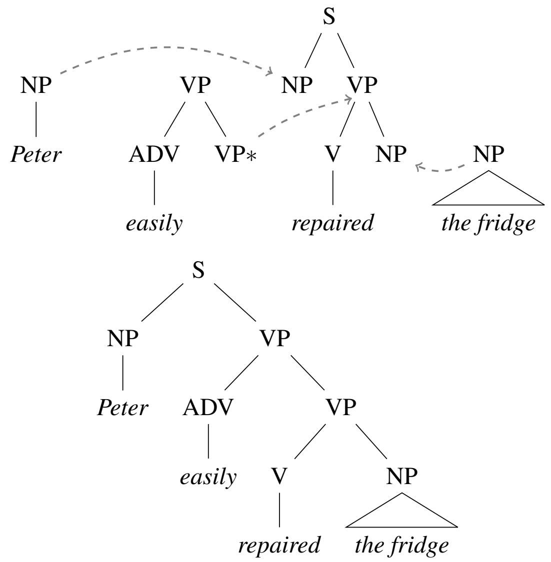

Figure 1: TAG derivation for the sentence *Peter easily repaired the fridge*.

One challenge lies in the complexity of Japanese scrambling, which can be shown with TAG. In this context, scrambling complexity refers to the generative power of the used TAG framework. In connection to this, the interesting question arrises how similar well-known scrambling languages such as German, Japanese, or Russian are and if they can be assigned to different complexity classes. I will argue that Japanese can be captured satisfyingly by TL-MCTAG, in contrast to German or presumably Russian. Becker et al. (1992) show that German LDS cannot be captured by TL-MCTAG and suggests a more powerful variant of TAG. Japanese allows LDS, as well, but it turns out to be more restricted. Russian, on the contrary, seems even less restrictive on scrambling compared to German (Sekerina, 2003).

In this work, I intend to provide LTAG analyses for scrambling in Japanese. As a result I will show, that LTAG is not powerful enough, whereas TL-MCTAG provides the desirable derivational power. Derivational power takes into account as to how derived structures are formed, instead of just regarding the derived structures themselves.

2 Word Order in Japanese

Tsujimura (2000) proposes six rules that restrict Japanese free word order. Firstly, as a strict head-final language, the verb cannot be scrambled and every constituent precedes the verb. (2a) shows the canonical order, while (3) opposes the first restriction.

(3) * Kaito-ga itta$_i$ [kinō Tarō-ga Ginza-de sushi-to-sashimi-o tabeta] to Hanako-ni t$_i$.

Secondly, the noun phrase and its corresponding particle are always considered a constituent and cannot be reordered separately (4). Also, particles always follow and assign the noun phrases, which is why constituents are represented as P(article) P(hrase)-leaves instead of NP-leaves. The examples (4)–(6) correspond to the glossing of (2a).

(4) * Kaito-ga [ga$_i$ kinō Tarō-$_i$ Ginza-de sushi-to-sashimi-o tabeta] to Hanako-ni itta.

Thirdly, the connective word *to* joins two or more NP to one constituent, so that it becomes ungrammatical to scramble only one of those conjoined NPs as in (5).

(5) * Kaito-ga [kinō Tarō-ga sashimi-o$_i$ Ginza-de t$_i$-sushi-to tabeta] to Hanako-ni itta.

Fourthly, scrambling out of the embedded clause, or rather out of the domain of the governing verb, also has its limits. Subjects and modifier are bound to their head and cannot be scrambled out of the verbal phrase. Otherwise, it leads to ill-formed sentences such as (6). Scrambling constituents other than the subject or a modifier, however, is possible and leads to LDS, see again (2b).

(6) * Ginza-de$_i$ Kaito-ga [kinō Tarō-ga t$_i$ sushi-to-sashimi-o tabeta] to Hanako-ni itta.

(2b) Sushi-to-sashimi-o$_i$ Kaito-ga [kinō Tarō-ga Ginza-de t$_i$ tabeta] to Hanako-ni itta.

Fifthly, and this is connected to the head-final structure, only leftward movement is allowed, and thus, extraposition is ruled out. Finally, particle order may forbid scrambling, as well. For instance, a *ga-ni* particle order is free for scrambling, while a *ga-ga* order forbids LS of both constituents. The particle *ga* usually is a subject marker.

For LDS, two more conditions are required: complementizing and a governing verb of perception. In a variety of articles (Saito, 2012; Suzuki, 1994, among others) examining Japanese free word order, it is stated that the *to*-complementizer is prominent to permit scrambling out of sentence boundaries. The *to*-complementizer marks direct and indirect speech, and thus will be denoted as a QUOTATION PARTICLE (PQ) POS-TAG. In fact, there exist more complementizers which permit LDS, i.e., the FORMAL NOUNS (NF[1]) *koto*, and *no*, or *yō* and *ka*. Together with these complementizers, verbs like *iu* ('say'), *omou* ('think'), *shiru* ('know'), or *kangaeru* ('reason') allow for LDS.

Furthermore, complementizing can be divided into two groups of precedence patterns. In both structures the complementizing (COMP) part is preceded directly by the verb of the embedded clause (V2). Then, the governing verb (V1) succeeds the complementizer, but also allows constituents, a dative case for instance, in between. Yet, *to* or *ka* complementizer do not demand any particles (P), as suggested in (7). A NF, however, requires an immediately succeeding particle, since it nominalizes the relative clause. This precedence structure is shown in (8). Strict precedence is marked by $>$, whereas $\gg$ is non-strict precedence.

(7) $\text{V2} > \text{COMP} \gg \text{V1}$

(8) $\text{V2} > \text{NF} > \text{P} \gg \text{V1}$

(7) and (8) are crucial for the TAG analysis proposed in Section 3.2, insofar that V1, V2, COMP, NF, and P will appear as leaves in the elementary trees.

3 TAG Modelling

3.1 Underlying Linguistic Principles

TAG is a mathematical formalism in the first place. It lacks the linguistic interpretation, e.g., principled constraints on the shape of elementary trees of the nodes and syntactic structure. In this article, I use the valency principle according to Frank (2002) and Lichte (to appear, Section 5.3) as linguistic interpretation. In short, the lexical anchor represents the valency head or carrier and the nonterminal leaf nodes the valency roles. Modification, on the other hand, is factored away into a separate elementary tree (such as *easily* in Fig. 1). Still, as Frank (2002, 22) points out as the Fundamental TAG Hypothesis, every syntactic dependency, such as valency relation, is expressed locally within an elementary tree. The valency head co-occures with its arguments, which are nonterminal leaves (such as both NP-nodes of *repaired* in Fig. 1). Functional trees do not realize any valency and thus, the valency principle does not apply (Lichte, to appear, Section 5.3).

3.2 Scrambling with LTAG

Each instance of LS can be easily linearized in one elementary tree, as can be seen from the elementary trees in Fig. 2. The valency carrier (head) *ageta* ('gave') has three valency roles realized in one elementary tree. For each linearization the PP leaf nodes take constituents with the fitting particles. The constituents, which substitute into the PP leaf nodes, can be realized at only that position.

Realizing LDS with LTAG needs a more elaborate approach, since scrambling outside the domain of a head needs a particular sentence structure. This condition is realized with the PQ and NF elementary trees in Fig. 3, and in accordance to the precedence relations proposed in (7) and (8). The first requirement to these auxiliary trees is to permit embedding of further sentences. Verbs like *omou* ('think') would need a NF node, which enables embedding, while verbs like *iu* ('say') need a quotation particle *to*. Furthermore, a NF itself needs a succeeding particle, while PQ does not, resulting in slightly different auxiliary trees, as shown in Fig. 3. The initial tree and auxiliary tree of *itta* ('said') are used for the derivation in Fig. 4.

Note that the number of PP-leaves may differ as they represent valency roles. The VP-footnode adjoins to the initial tree, which is supposed to be the embedded tree after derivation. In addition, and on the basis of FTAG, VP nodes of the elementary trees carry feature structures with a Boolean SUBJ $+$ attribute, in contrast to the VP-foot node, which is enriched with SUBJ $-$. The PP that hosts a subject passes this information on to the VP-child of his VP-sister. This mechanism makes sure that no auxiliary tree with a

[1]This term was introduced in the Japanese treebank of the VERBMOBIL project (Kawata and Bartels, 2000, 28). The semantic content of formal nouns is empty and they are used to form nominal structures together with other expressions.

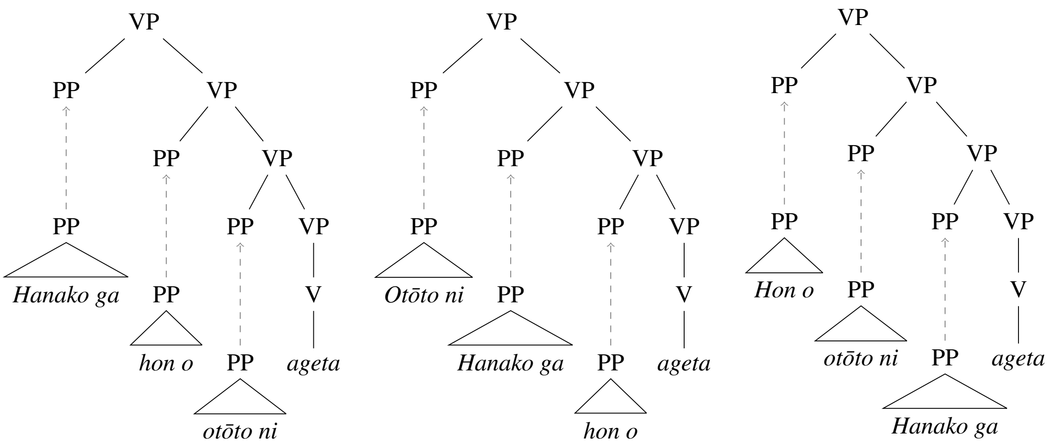

Figure 2: Elementary trees for the local scrambing of arguments of *ageta* ('give'). They are needed to derive the sentences in (1).

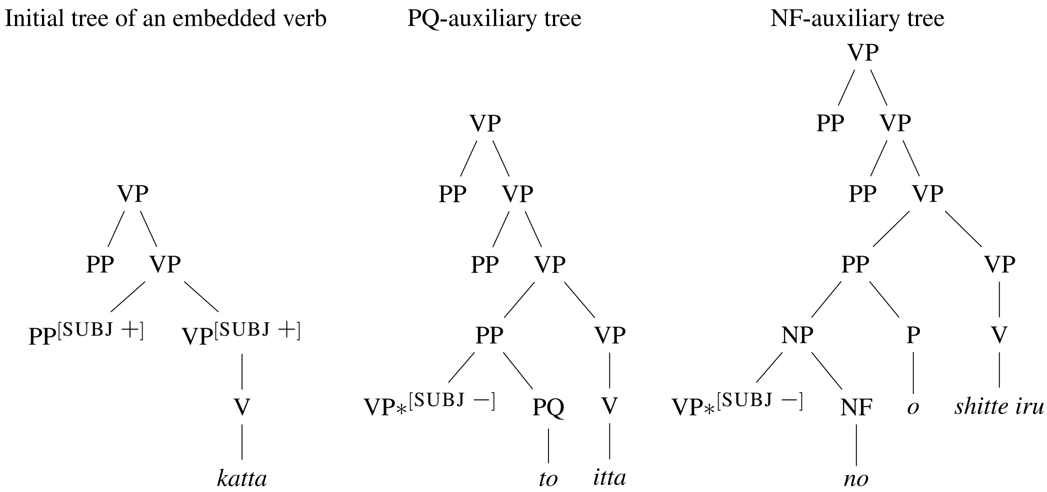

Figure 3: Elementary trees: initial tree, PQ- and NF-auxuliary trees.

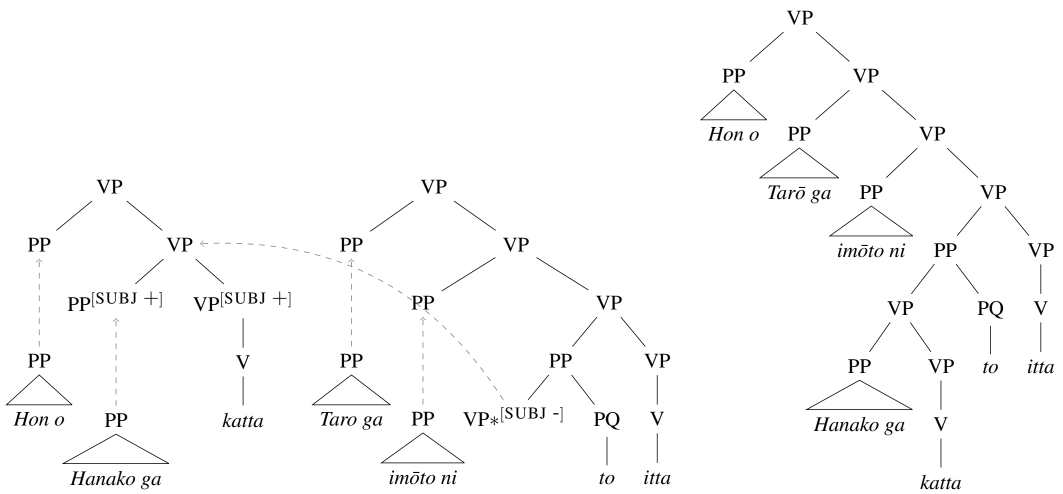

Figure 4: TAG derivation of (9), where *hon o* ('book', accusative) undergoes LDS.

SUBJ −-marked footnode can adjoin to a (partial) tree, which is dominated by a leaf with a SUBJ +-marking, and thus, restrict long-distance scrambling. An LTAG derivation of sentence (9), where *hon o* ('book', accusative) is the long distance scrambled constituent, is shown in Fig. 4. The sentence is glossed with the corresponding string schema (Becker et al., 1991), where $N1$ and $N2$ are valency roles of $V1$ and $V2$, respectively.

(9) Hon-o Tarō-ga imōto-ni Hanako-ga katta
N2 N1 N1 N2 V2
to itta.
V1
'Taro told to the little sister that it is the book, which Hanako bought.'

However, LTAG is not sufficient to fully cover LDS in Japanese. The auxiliary tree of Fig. 4, which has the string schema $N1$ $N1$ $V1$, where N corresponds to the valency role of the head V, adjoins to the initial tree ($N2$ $N2$ $V2$), resulting in the string schema $N2$ $N1$ $N1$ $N2$ $V2$ $V1$.[2] The derivational power (Becker et al., 1991) of LTAG meets its limits with the sentence structure $N1$ $N2$ $N1$ $N2$ $V2$ $V1$. From the derivational point of view, there is no possibility to generate such a structure with tree templates as in Fig. 3. Prospective new trees (see Fig. 5)[3] would need additional inner nodes for adjunction and also the relation between auxiliary tree and initial tree would be changed, since the footnode is in the primary initial tree. Thus, the tree modelling would be inconsistent, and additional inner nodes in the initial tree for the sake of adjunction would become necessary. On the other hand, auxiliary trees would lack non-terminal leaves. Hence, the valency principle would be violated. Also, the projection of the complementizer, which has to be realized as an auxiliary tree, contradicts the valency principle by being realized in the initial tree.

3.3 TL-MCTAG: Gaining more Derivational Power

Chen-Main and Joshi (2014) propose TREE-LOCAL MULTI-COMPONENT TAG (TL-MCTAG), which is equal to TAG in generative power, but more powerful in derivational terms,

[2]Note that the sentence structure $N2$ $N1$ $N2$ $V2$ $N1$ $V1$ also lies within the derivational power of LTAG.

[3]The meaning of this sentence is slightly different: 'Taro told the little sister, that it is a book which Hanako bought'

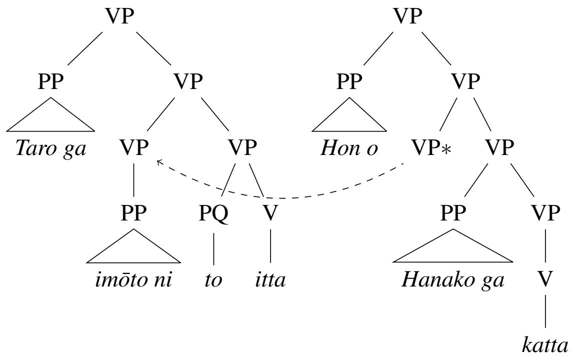

Figure 5: Contradiction of the Valency Principles: new trees for deriving the string schema $N1$ $N2$ $N1$ $N2$ $V2$ $V1$.

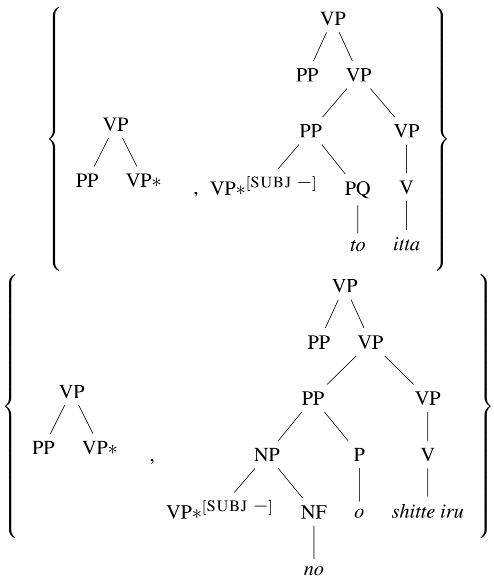

Figure 6: MC-Tags of PQ- and NF-trees from Fig. 3.

for ill-nested dependency structures or gap degree > 1. TL-MCTAG, as proposed in Fig. 6, permits to integrate the auxiliary trees of Fig. 3 as MC-sets. Doing so, string schemas such as $N1$ $N2$ $N1$ $N2$ $V2$ $V1$ can be generated without contradicting the valency principles other than the TAG analysis in Fig. 5. Fig. 7 proposes the derivation of the string schema in Fig. 5 with TL-MCTAG.

Furthermore, string schemas, which could be problematic for TL-MCTAG, do not appear in Japanese. For instance, since extraposition is ruled out, there is no possibility of a valency role preceding the head. Also, even if constituents are scrambled according to the restrictions in Section 2, LDS of more than one constituent results in un-

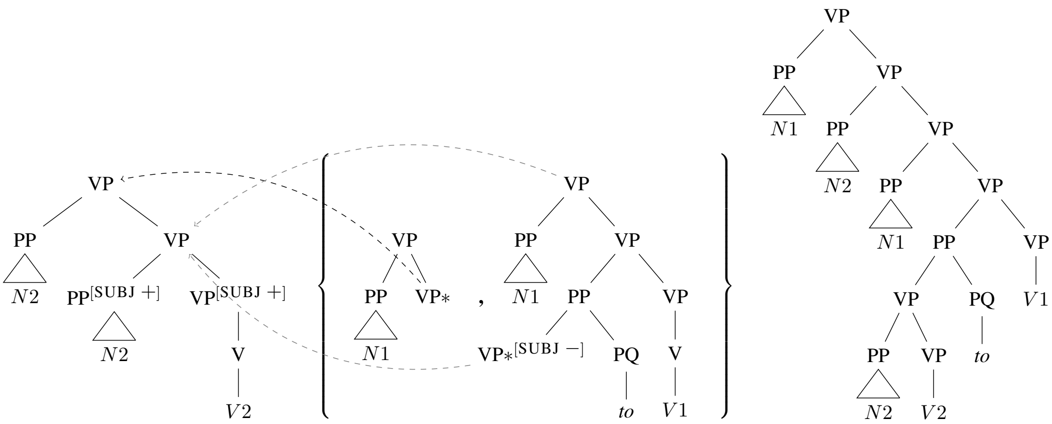

Figure 7: TL-MCTAG derivation of the string schema $N1\ N2\ N1\ N2\ V2\ V1$.

grammatical sentences. The word order in (10a), with the scrambled *Hisarya-e*, is a grammatical sentence in Japanese, while the additional scrambling of *hon-o* in (10b) results in an ungrammatical sentence.[4]

(10) a. Hisarya-e Tarō-ga Ichirō-ni
Hissar-to Tarō-NOM Ichirō-DAT
Hanako-ga hon-o okutta
Hanako-NOM book-ACC send.PST
to itta
COMP say.PST
'Taro told Ichiro, that it is Hissar, where Hanako send the book.'

b. * Hisarya-e Tarō-ga hon-o Ichirō-ni
N2 N1 N2 N1
Hanako-ga okutta to itta
N2 V2 V1

4 Complexity Comparison

With TAG it is possible to compare language complexity, considering scrambling complexity, in particular. As discussed in this paper, a fully sufficient TAG variant for Japanese, in the terms of generative and derivational power, is TL-MCTAG. Becker et al. (1992) show, however, that TL-MCTAG does not suffice for German, since scrambling (LS and LDS) is unbound to distance and number of dependencies. Russian, on the other hand, can only be vaguely categorized, due to a lack of literature that discuss Russian LDS exhaustively. Glushan (2006) shows that Russian LDS has similarities with Japanese LDS, but is less restricted. For instance, in Russian it is permitted in some cases to dislocate subjects and modifier out of embedded clauses. Yet, Glushan (2006) points out that scrambling is also more restricted than assumed in the literature, since she could list a number of cases, where LDS is either unrestricted or clearly restricted. Additionally, she argues that Russian scrambling can be successive cyclic. It is unclear, though, whether 'doubly unbounded' constructions are possible, similar to German. Thus, Japanese scrambling is less complex than German, since TL-MCTAG is sufficient for Japanese but not for German anymore. Russian appears to be more complex than Japanese, but it remains unclear if the complexity is lower, equal, or even higher than German.

5 Conclusion

In this paper I have shown that scrambling in Japanese, even though being very flexible within local domains, underlies considerable constraints when it becomes non-local. Eventually, these constraints require some derivational power (in terms of string schemata) that is already available in TL-MCTAG. This result is in sharp contrast to TAG-approaches to scrambling in other languages, notably German (Becker et al., 1992), where much more powerful extensions of TAG are necessary. Another language of this sort seems to be Russian (Sekerina, 2003). It therefore seems that scrambling cross-linguistically falls into different complexity classes, which can be neatly characterized within the TAG-framework.

References

John Frederick Bailyn. 2002. Scrambling to reduce scrambling. *Glot International*, 6(4):83–90.

[4] I'm grateful to Mamoru Saito for helping me with this sort of sentences

Tilman Becker, Aravind K. Joshi, and Owen Rambow. 1991. Long-distance scrambling and Tree Adjoining Grammars. In Jürgen Kunze and Dorothee Reimann, editors, *EACL '91 Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, pages 21–26.

Tilman Becker, Owen Rambow, and Michael Niv. 1992. The Derivationel Generative Power of Formal Systems or Scrambling is beyond LCFRS. Technical Report IRCS-92-38, Institute for Research in Cognitive Science, University of Pennsylvania.

Joan Chen-Main and Aravind K. Joshi. 2014. A dependency perspective on the adequacy of Tree Local Multi-Component Tree Adjoining Grammar. *J. Log. Comput.*, 24(5):989–1022.

Robert Frank. 2002. *Phrase structure composition and syntactic dependencies*, volume 38 of *Current studies in linguistics*. MIT Press, Cambridge and Mass.

Zhana Glushan. 2006. *Japanese Style Scrambling Russian: Myth and Reality*. Ph.D. thesis, University of Tromsoe.

Aravind K. Joshi and Yves Schabes. 1997. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 3, pages 69–124. Springer, Berlin, New York.

Yasuhiro Kawata and Julia Bartels. 2000. Stylebook for the Japanese treebank in Verbmobil.

Timm Lichte. to appear. *Syntax und Valenz: Zur Modellierung kohärenter und elliptischer Strukturen mit Baumadjunktionsgrammatiken*. Language Science Press, Berlin.

Mamoru Saito. 2012. Sentence types and the Japanese right periphery. In G. Grewendorf and T. E. Zimmermann, editors, *Discourse and grammar*, volume 112 of *Studies in generative grammar*, pages 147–176. de Gruyter Mouton, Boston and Berlin.

Irina Sekerina. 2003. Scrambling and processing: Dependencies, complexity, and constraints. In Simin Karimi, editor, *Word order and scrambling*, volume 4 of *Explaining linguistics*, pages 301–324. Blackwell Pub., Malden and Mass.

Satoko Suzuki. 1994. That a fact? Reevaluation of the relationship between factivity and complementizer choice in Japanese: Proceedings of the twentieth annual meeting of the berkeley linguistics society: General session dedicated to the contributions of charles j. fillmore. *The Berkeley Linguistics Society*, 20(1):521–531.

Natsuko Tsujimura. 2000. *An introduction to Japanese linguistics*, volume 10 of *Blackwell textbooks in linguistics*. Blackwell, Cambridge and Mass, repr. edition.

K. Vijay-Shanker and Aravind K. Joshi. 1988. Feature structure based Tree Adjoining Grammars. In *In proceedings of COLING*, pages 714–719.

David J. Weir. 1988. *Characterizing mildly context-sensitive grammar formalisms*. Ph.D. thesis, University of Pennsylvania.