AN ANNOTATED CORPUS IN JAPANESE USING TESNIÈRE’S STRUCTURAL SYNTAX

Yves LEPAGE§, ANDO Shin-Ichi†, AKAMINE Susumu‡, IIDA Hitoshi§
§ATR Interpreting Telecommunications Research Labs, Kyoto, Japan
{lepage, iida}@itl.atr.co.jp
NEC, †C&C Media and ‡Human Media Research Labs, Kanagawa, Japan
ando@ccm.cl.nec.co.jp, akamine@hml.cl.nec.co.jp

INTRODUCTION

Tesnière’s attention to covering a maximal number of syntactic phenomena explains the impressive number of languages – “timem hominem unius linguae” – cited in the Eléments de syntaxe structurale. Although Japanese is correctly classified as a strongly centripetal language according to linear survey (relevé linéaire, p. 33), no examples of Japanese are cited. Consequently, we have endeavored to apply Tesnière’s ideas to Japanese by manually constructing the linguistic structures for more than six thousand sentences of a corpus of hotel reservation conversations.

In fact, Tesnière’s grammatical ideas, and among them, the most original ones, fit well to Japanese as they give simple and insightful descriptions of some usually controversial grammatical phenomena (ergative constructions, na-adjectives).

After describing the different types and categories of words, we will focus on the three phenomena to which, according to Tesnière, all syntactical phenomena reduce: connection, junction and transference. From the representational point of view, we will introduce correspondence intervals to code which part of the surface text corresponds to which nodes or subtrees.

1 WORDS

We have taken the character (kana or kanji), which is the physical unit of a Japanese text, as the unit of measure of the length of a section of text. With the convention of starting at position 0, we locate any piece of text, and hence words, using an interval notation. Note that there is no word separator (or blank spaces) in Japanese. In the following sentence¹, the word 部屋 is located by the interval [3..5] and the word 飯えて by [6..9]. This notation will be used in correspondences (Section 2.2).

Could I get a room upstairs?

1.1 Species and Categories of Words

The differentiation between: content words, which are associated with a concept, and function words, which express syntactical information was not difficult to apply to Japanese.

1.1.1 Content Words

Some examples of content words include: 予約 (yoyaku, reservation), 逾期 (okureru, to be late), 高い (takai, expensive), 直接 (tyokusetu, directly). Tesnière distinguishes between two categories of content words: processes and substances, which are, for explanation purposes, usually exemplified by verbs and nouns, respectively, in Indo-European languages. This is also consistent with Japanese.

These two categories are in turn divided into: concrete and abstract categories, which opposes the concrete notion of processes and substances to their abstract attributes, and gives rise to the following categorisation for content words (see also (Starosta 88), Tesnière’s notations is shown in capitals).

|          | substances | processes |
|----------|------------|-----------|
| concrete | substantive | verbal    |
| abstract | adjectival  | adverbial |

¹Except when mentioned, examples are from the treebank.
It is to be noted that, in the case of Japanese, two categories of words are variable in relation to aspect and negation: abstract substances (A) and concrete processes (I), which are respectively (i-)adjectives and verbs in terms of Japanese grammars.

Now, some classes of words, which pose problems in Japanese grammar books written in English, such as the so-called na-adjectives (静か (sizuka, quiet)), and the Sino-Japanese nouns-verbs formed in conjunction with use of the Japanese verb する (suru, to do), can easily be categorised as nouns (O). This is consistent with what is taught in Japanese schools, (see Appendix B), their syntactical behaviour being perfectly described by transference (see Section 4).

1.1.2 Function Words

Grammatical tools, the role of which is to either make explicit, or change the category of a content word, or to define relationships between words, are called function words. These words will appear in extenso in structural representations.

In Japanese, many can be easily identified, such as, が (ga, 格助詞, nominative case post-particule), の (no, 連体助詞, genitive case post-particule), で (de, 接続助詞, equivalent to subordinate conjunction), か (ka, 疑助詞, end of interrogative sentence particle), する (suru, 助語尾, support verb for Sino-Japanese nouns), etc.

Of course, some function words can also be content words in a different context. For instance, the verb する (suru), is either the support verb for Sino-Japanese nouns, (a function word in that case), or the verb “to do” (a content word).

2 CONNECTION

Tesnière speaks of connection to describe the relations between words in a sentence in terms of their subordination relations. This concept includes predicate-argument or governor-modifier relations as well as predicate-circumstantial relations (Eléments, p. 14).

The study of sentences, which is the proper object of structural syntax is essentially the study of its structure, i.e. the hierarchy of its connections.

2.1 Tree Representation: Stemmas

Tesnière was the first to propose, in 1934, to systematically use graphical representations which he called stemmas, for representing this hierarchy (Tesnière 34). However, these stemmas were more than simple trees. Although, we will show that the introduction of correspondence makes it possible to encode Tesnière’s representations using just trees.

Basic connections are those which link concrete notions with their abstract attributes (Figure 1).

```
ヤスイ・ホテル = 快く
便宜に 帰る

便宜 = ホテル

進行ます

便宜 = 快く
帰る

便宜 = ホテル
便宜 = 快く
帰る

Figure 1: Basic connections.
```

By replacing content words with their class (O, I, A, E) “virtual” stemmas (on the right) can be derived from the “real” ones (on the left).

2.2 Correspondences

To explicitly indicate which word, or more specifically, especially in the case of Japanese, which chunk of text corresponds to which node in the stemma, we adopted the use of correspondences (Boitet and Zaharin 88).

We note two kinds of correspondence:

• words-to-node, and
• sentence parts-to-complete subtree, or substring-to-subtree.

Constraints Correspondences are noted by intervals, as introduced above, and are governed by three constraints (Lepage 94).

• global correspondence: an entire tree corresponds to an entire sentence;

He acknowledged that two Russian linguists used trees in 1930 to explain some syntactic phenomenon, but, unlike Tesnière, the use of trees was not pivotal in their explanations.
• inclusion: a subtree which is part of another subtree $T$, must correspond to a substring in the substring corresponding to $T$;
• membership: a node in a subtree $T$, must correspond to words members of the substring corresponding to $T$.

Breakfast is not included.

In Figure 2, on each node of the stemma, two intervals stand for the words-node and the substring-subtree correspondences, in that order. The entire sentence extends from 0 to 11, as indicated by the root. This root is a verb, denoted as I, and is located in position 7 to 11: ならます (narimasu). Similarly, the node labelled ト (hi) corresponds as a word to the case-maker へ, which extends from 6 to 7 in the sentence. The entire subtree dominated by the node corresponds to the phrase 別料金 に (beturyoukin ni) which extends from 3 to 7.

Discontinuous Intervals Discontinuous intervals are possible. In Figure 3, the deverbal noun 願い (negai, request) from 願う (negau, to ask for) takes an accusative argument extending from 0 to 4, お名前を (o+namae wo, your name). Because the honorific prefix お (o+) can only be applied to a noun, obtained by attaching the suffix い (+i) (transference, see Section 4), the subtree dominated by the verbal root corresponds to a non-connex substring [0.4]+[5.6] in the surface form: お名前を...願います 10.

What is your name, please?

2.3 Predicate-Argument Structures

Free-Order – Subject A main feature of dependency structures, to which Tesnière's representations pertain, is that they do not provide any preferred position to the subject (see Fourquet's foreword to (Gréciombo and Schumacher 96), and (Zemba 76), p. 393, for a discussion). This corresponds particularly well with our data because the free ordering of case-marked phrases (not words) is a property of Japanese, which makes dependency grammars more adequate in its description. For exam-

Figure 2: A sentence and its associated stemma.

Figure 3: A case of a discontinuous interval.
pie, the two following propositions are equally valid, where location and subject have been exchanged.

六人 が 一部屋 に 入れる
rokuinin ga hitoheya ni ireru
‘6-people’ NOM ‘1-room’ LOC ‘can-enter’

a room that can accommodate 6 people

Omission Moreover, in Japanese, the omission of any of the case-marked phrases is possible. One can perfectly imagine a situation where a traveler first announces that he is in a group of 6 people, and then merely utters the following sentence:

一部屋 に 六人 が 入れる...
hitoheya ni rokuinin ga ireru...
‘1-room’ LOC ‘6-people’ NOM ‘can-enter’

This sentence has no subject, and yet it is unambiguously understood as a request for a room which can accommodate 6 people altogether.

Ergative Constructions Moreover, the search for the “real subject”, as opposed to the syntactical subject, is meaningless in dependency representations of ergative constructions. Such constructions exist in Japanese with a range of adjectives, such as, 欲しい (hosii) (20 occurrences in our corpus), or verbal forms in -たい (taii) (around 310 occurrences in the corpus), or the so-called “passive” or “medio-passive” verbs, such as, 見える (mieru, cf. Fr. se voir).

verb, has to be postulated for Japanese, because...it is so in English.

However, the ergative case does not exist in Japanese, and it would be difficult to call Japanese an ergative language (see Mel’čuk 88), p. 259–253, for definitions concerning ergativity).

Auxiliary Verbs In an original and interesting discussion, Tesnière advocates that the subject and the object of a French passé composé of a transitive verb, do not both link to the past participle. He shows that some clues indicate that the subject links to the auxiliary, while the object should be linked with the past participle. Similar analysis seems particularly well adapted to some Japanese constructions too, not because of the agreement in gender-number, but because of case semantics.

For instance, in the following sentence, the subject, postal code, cannot be considered the subject of the verb, to write.5

auxiliary

The postal code is written (e.g. on an envelope)

However, changing the auxiliary, ある (aru) into いる (iru) implies a change in the case of postal code.

The postal code is being written (e.g. by Lucien) = Somebody is writing the postal code.

This convinced us to adopt Tesnière’s analysis, where the subject is linked with the auxiliary (Figure 5).

Figure 4: An ergative construction.
3 JUNCTION

Junction gathers the facts of coordination, and factorisation. Junction words in Japanese include words such as と (to, and for nouns), や (ya, or for nouns), し (si, or for verbs), けど (kedo, but). We propose to represent them with one node bearing a special label: we prefix and suffix by - the function word. Accordingly, we can easily represent cap junctions as in Figure 6.

On the other hand, in cup cases, the same dependent shares several governors. A tree can be “factored” by using a special node, V, bearing the same correspondences as its root. Figure 7 is a slightly modified corpus sentence.

4 TRANSFERENCE

Transference, in essence, consists in transferring to a content word of a given category the function or role of another category. According to Tesnière, it is precisely this transference which allows a speaker of any language to never be stopped by the fact that a needed concept does not fit, by category, into the role required at a given point in an utterance.

Transferee: Transference applies to a content word, called the transferee. It is performed by a transferer, which may be:

- a function word の (no, of), に (ni, to), す る (suru, to do), etc.

Here, we follow the recommendation of Tesnière himself to render the French word translation with this English term especially coined for the meaning here.
some morphological device + < (ku, adver-
bial form of adjectives), + て (te, pending
form of verbs), etc.

no mark at all (the so-called “relatives” of
Japanese are in fact transferences: a verb
is transferred into an adjective without any
marker). In this case, we indicate the trans-
fer node by が.

As a result of transference, the category of
the content word has been transformed into an-
other category so that it can play the role of the
resulting category. For instance:

ホテル:O → ホテルの:A
hote ru  hote ru no
hotel   of the hotel

Representation Depending on the position
of the transferer, left and right transferences
have to be distinguished. In Japanese, the
transferer is predominantly on the right of the
transferee. We represent the transference with
the help of a 3-node subtree to render Tesnière’s
capital T notation:

- the mother node bears the target category,
  followed (or preceded) by T if the transferer
  is on the right of the transferee in the sen-
tence, (usual case in Japanese), or on the
  left;
- the left (or right) daughter bears the trans-
feree, represented by its category;
- the other daughter bears the transferer, i.e.
  the function word in extenso.

A mother node does not correspond to any
word in the surface text so it bears an empty
interval (denoted as [n_n], with any n). How-
ever, as the root of a subtree, it represents the
sum of the intervals of all its subtrees.

Na-Adjectives A class of Sino-Japanese
nouns exists in Japanese, extended in contem-
poranean Japanese by a full range of English-
Japanese nouns (Sells 96) (ユニークな (yuniku-
na, unique), フレッシュな (huressyu-na, fresh)),
which could be semantically interpreted as ad-
jectives, but follow a specific syntactical be-
haviour, different from standard adjectives end-
ing in い (i) (Appendix C). They are the so-
called na-adjectives in Japanese grammar books

written in English, although in Japanese termi-
nology they are described as noun-adjectives.

In attributive positions, these words require
a special function word, な (na). We analysed
な as a transferer of nouns (O) into adjectives
(A). This view meets that of (Kuwae 89), vol 1,
p. 185, who considers that, “だ (da) is the only
variable word in Japanese for which there exists
determinative form, な (na), distinct from the
conclusive form.”

CONCLUSION
We have presented a tree-bank of 6553 sentences
of Japanese conversations in the domain of ho-
tel reservations, which uses Tesnière’s structural
syntax framework. Correspondences between
surface texts and trees are ensured by means
of intervals.

It has long been felt in the NLP Japanese
community that a dependency approach fits
well to the description of Japanese. The
privileged place for the subject in con-
stituency descriptions generates artificial prob-
lems, whereas, dependency allows simple and
direct description of phenomena like, for in-
stance, ergative constructions.

Moreover, Tesnière’s original ideas give a
clear insight to some area. For instance, the
attachment of arguments under auxiliaries bet-
ter renders case semantics. Also, transference
permits a simple analysis of “na-adjectives”,
which respects the feeling of native speakers of
Japanese.

Figure 8: Representation of transference.
A Grammatical Labels Used in Glosses

| symbol | particle or example | example |
|--------|---------------------|---------|
| TOP    | は (ha)             | topicalisation |
| NOM    | が (ga)             | nominative   |
| ACC    | を (wo)             | accusative   |
| LOC    | で (de)             | locative     |
| HON    | お+ (go+)           | honorific    |

B Structural Syntax Categories and Japanese School Grammar Classes

| symbol | class | example |
|--------|-------|---------|
| 0      | 名詞 | 名詞   |
| A      | 形容詞 | 形容詞 |
| I      | 動詞 | 動詞 |
| E      | 副詞 | 副詞 |

C I-Adjectives and "Na-Adjectives"

| polite     | predicative | familiar | attributive |
|------------|-------------|----------|-------------|
| takai desu | 高い        | 高い     | 高い         |
| sizuka desu| 静か         | 静か     | 静か         |

References

Christian Boitet and Zaharin Yusoff
Representation trees and string-tree correspondences
Proceedings of COLING-88, Budapest, 1988, pp 59-64.

Gunji Takao
Japanese Phrase Structure Grammar
D. Reidel Publishing Company, 1987.

Gertrud Gréciando und Helmut Schumacher
(Herausgegeben von)
Lucien Tesnière - Syntaxe structurale et opérations mentales
Max Niemeyer Verlag, Tübingen, 1996.

Kunio Kuwae
Cours de japonais
vol. 1 & 2, L'Asiathèque, Paris, 1989.

Yves Lepage
Texts and Structures – Pattern-matching and Distances
ATR report TR-IT-0049, Kyoto, March 1994.

Igor A. Mel'čuk
Dependency Syntax: Theory and Practice
State University of New York Press, 1988.

Peter Sells
What Happens When A Word is Borrowed
handout of communication, ATR-ITL, July 11th, 1996.

Stanley Starosta
The Case for Lexicase
Pinter Publishers, London and New York, 1988.

Lucien Tesnière
Comment construire une syntaxe
Bulletin de la Faculté des Lettres de Strasbourg, 12e année, n° 7, mai-juin 1934, pp. 219-229.

Lucien Tesnière
Eléments de syntaxe structurale
Klincksieck, Paris, 1959.

Jean Marie Zemb
Vergleichende Grammatik Französisch-Deutsch
Comparaison de deux systèmes - Teil 1
Dudenverlag, 1978.