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 – "*timeo 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 sentence[1], the word 部屋 is located by the interval [3_5] and the word 変えて by [6_9]. This notation will be used in correspondences (Section 2.2).

$_{0}$上$_{1}$階$_{2}$に$_{3}$部$_{4}$屋$_{5}$を$_{6}$変$_{7}$え$_{8}$て$_{9}$く$_{10}$だ$_{11}$さ$_{12}$い$_{13}$。

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), 遅れる (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 O	verbal I
abstract	adjectival A	adverbial E

[1] Except 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 prefectly 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-particle), ので (node, 接続助詞, equivalent to subordinate conjunction), か (ka, 終助詞, end of interrogative sentence particle), する (suru, サ変語尾, 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[2] 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).

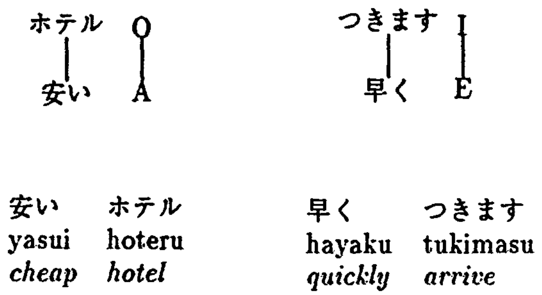


Figure 1: Basic connections.

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:

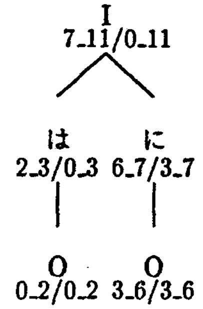
- words-to-node, and
- sentence parts-to-complete subtree, or substring-to-subtree.

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;

[2] 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.



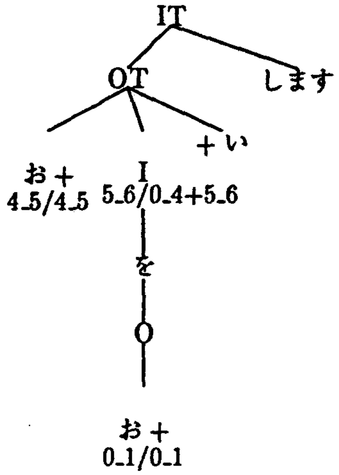
$_0$朝食$_2$	は$_3$	別料金$_6$	に$_7$
tyousyoku	ha	beturyoukin	ni
'breakfast'	TOP	'extra charge'	LOC
なります$_{11}$			
narimasu			
'become'			

Breakfast is not included.

Figure 2: A sentence and its associated stemma.

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[3] 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: なります (narimasu). Similarly, the node labelled に (ni) corresponds as a word to the case-maker に, 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 deverbative 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.



お$_1$	名前$_3$	を$_4$	お$_5$	願い$_7$	します$_{10}$
o	namae	wo	o	negai	simasu
HON	'name'	ACC	HON	'request'	'do'

What is your name, please?

Figure 3: A case of a discontinuous interval.

2.3 Predicate-Argument Structures

Free-Order – Subject A main feature of dependency structures, to which Tesnière's representations pertains, is that they do not provide any preferred position to the subject (see Fourquet's foreword to (Gréciano and Schumacher 96), and (Zemb 78), 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[4]. For exam-

[3] Refer to Table A in Appendix for notations used in glosses.

[4] (Mel'čuk 88) and (Starosta 88), among others have already commented that constituency structures are English-oriented representations into which some linguists try desperately to cast other languages. An illustration is (Gunji 87). After a ten-page discussion, and despite an honest acknowledgment that there is absolutely no basis for this, he draws the conclusion that a preferred position for the subject, as a left sister of the

ple, the two following propositions are equally valid, where location and subject have been exchanged.

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

a room that can accommodate 6 people

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

a room that can accommodate 6 people

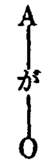
Omission Moreover, in Japanese, the omission of any of the case-marked phrases is possible. One can perfectly imagine a situation where a traveler first announces that he is in a group of 6 people, and then merely utters the following sentence:

一部屋	に	入れる
hitoheya	ni	ireru
'one-room'	LOC	'can-enter'

a room that can accommodate 6 people

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[5] with a range of adjectives, such as, 欲しい (hosii) (20 occurrences in our corpus), or verbal forms in - たい (tai) (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).



水中眼鏡	が	欲しい...
suityuumegane	ga	hosii
'goggles'	NOM	'want'

I want goggles

Figure 4: An ergative construction.

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*[6].

郵便番号	が	書いて	ある
yuubinbangou	ga	kaite	aru
'postal code'	NOM	'write'	'is'

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.*

郵便番号	を	書いて	いる
yuubinbangou	wo	kaite	iru
'postal code'	ACC	'write'	'is'

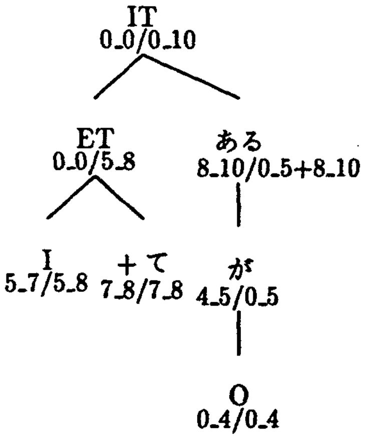
The postal code is being written (e.g. *by Lucien) = Somebody is writing the postal code.*

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

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

[5] 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. 250-253, for definitions concerning ergativity).

[6] 書いて (kaite) is a non-conclusive, pending, form of the verb 書く (kaku), which is translated in English by "writing" or "written" according to the context.



$_{0}$郵便番号$_{4}$	が$_{5}$	書いて$_{8}$	ある$_{10}$
yuubinbangou	ga	kaite	aru
'postal code'	NOM	'write'	'is'

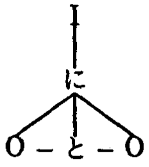
The postal code is written

Figure 5: Auxiliary dependency.

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.

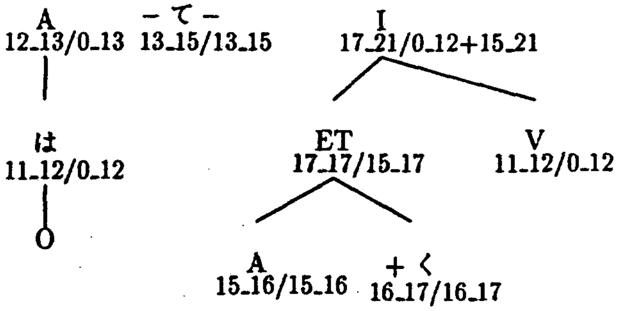


一階	と	地下一階	に
ikkai	to	tikaikkai	ni
'first-floor'	'and'	'basement floor'	LOC

On the first and basement floors

Figure 6: A case of a cap junction.

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.



$_{0}$国際エキスプレスメール$_{11}$	は$_{12}$	安くて$_{15}$
kokusai ekisupuresu meeru	ha	yasukute
'International express mail'	TOP	'cheap-and'

早く$_{17}$	つきます$_{21}$
hayaku	tukimasu
'quickly'	'arrive'

International express mail is cheap, and it arrives quickly

Figure 7: A case of a cup junction.

Because of junctions, a structure representing a sentence may be a forest. This is a significant difference to constituency representations, but conforms with Tesnière's description (*e.g.* p. 649). Figure 7 is such an example.

4 TRANSFERENCE

Transference[7], 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.

Transferer 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.*

[7] 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 + く (ku, adverbial 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 transferer node by ∅.

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

ホテル:O	→	ホテルの:A
hoteru		hoteru 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 sentence, (usual case in Japanese), or on the left;
- the left (or right) daughter bears the transferee, 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*). However, 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 contemporanean Japanese by a full range of English-Japanese nouns (Sells 96) (ユニークな (yuniiku-na, unique), フレッシュな (huressyu-na, fresh)), which could be semantically interpreted as adjectives, but follow a specific syntactical behaviour, different from standard adjectives ending in い (i) (Appendix C). They are the so-called na-adjectives in Japanese grammar books written in English, although in Japanese terminology they are described as noun-adjectives.

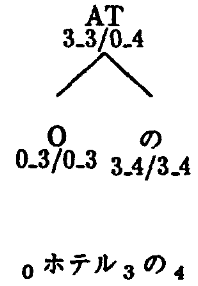


Figure 8: Representation of transference.

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 a 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 hotel 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 constituency descriptions generates artificial problems, whereas, dependency allows simple and direct description of phenomena like, for instance, ergative constructions.

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

A Grammatical Labels Used in Glosses

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

B Structural Syntax Categories and Japanese School Grammar Classes

symbol	class	example
O	名詞 nouns	ホテル（普通名詞） 予約（サ変名詞） 静か（形容名詞） 十九万ウォン（数詞） 伊藤（固有名詞）
A	形容詞 adj.	高い（形容詞） よろしい（形容詞） その（連体詞）
I	動詞 verbs	遅れる なります（+ます） 来ました（+た）
E	副詞 adv.	直接（副詞） はい（感動詞）

C I-Adjectives and "Na-Adjectives"

predicative		attributive
polite	familiar	
高いです takai desu *it is expensive*	高い takai *(it's) expensive*	高い部屋です takai heya desu *it is an expensive room*
静かです sizuka desu *it is quiet*	静かだ sizuka da *(it's) quiet*	静かな部屋です sizuka na heya desu *it is a quiet room*

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

岩淵匡, 桜井光昭, 武部良明, 森田良行 (共編)
日本文法用語辞典, 三省堂, 1989.

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.

益岡隆志 & 田窪行則
基礎日本語文法, くろしお出版, 1989.

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.