自然言語処理 134－19
（1999. 11. 26）

日本語における動詞交替の種類と頻度の予備調査

Timothy Baldwin, 徳永 健伸, 田中 穂積

東京工業大学 情報理工学研究科

〒152-8552 東京都目黒区大岡山 2-12-1

{tim,take,tanaka}@cl.cs.titech.ac.jp

概要

本研究は，日本語語彙体系の結合価辞書を解析し，日本語における動詞交替の種類と頻度を定量的に分析することを目的とする．本論文では，交替によって選択制約が変わらないものと仮定し，格スロット間の選択制約を比較して動詞交替を検出する．検出手法としては，(1) 選択制約の単純なマッチング，(2) バッキング・オフを許した選択制約のマッチング，そして (3) バッキング・オフを許した「意味密度」，の3種類の解析法を提案する．

Preliminary analysis of the range and frequency of Japanese verb alternations

Timothy Baldwin, Takenobu Tokunaga and Hozumi Tanaka

Dep. of Computer Science, Tokyo Institute of Technology

2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552

{tim,take,tanaka}@cl.cs.titech.ac.jp

Abstract

We set out to empirically identify the range and frequency of basic verb alternation types in Japanese, through analysis of the Goi-Taikei Japanese pattern-based valency dictionary. This is achieved through comparison of the selectional preference case marker annotation on corresponding case slots, based on the assumption that selectional preferences are preserved under alternation. Three separate extraction methods are considered, founded around: (1) simple match of selectional restrictions; (2) selectional restriction matching, with recourse to penalised backing-off; and (3) semantic density, again with recourse to backing-off.

1 Introduction

This is an attempt to document the full range of verb alternations in Japanese from analysis of structural regularities between entries in a valency dictionary, and determine the degree of permeation of each individual alternation type. Various techniques are analysed for extracting alternations, focusing on full or partial preservation of the selectional restrictions describing each case slot, and the degree of restriction of the selectional restrictions. These methods operate around the assumption that selectional restrictions are essentially unchanged under alternation.[1]

Verb alternations have been the target of considerable research within linguistic circles, in terms of describing the range of alternations for a particular language (e.g. (Levin, 1993) for English and (Fukui et al., 1985) for Japanese), applying alternations to model verbal semantics (e.g. (Levin, 1993; Hale and Keyser, 1987)), or analysing alternating potential and interpretational differences between verbs in different clause contexts (e.g. (Goldberg, 1995; Wierzbicka, 1988)). More recently, alternations have been the object of interest within the natural language processing community, for use in lexical selection in natural language generation (Dorr and Olsen, 1996; Jing and McKeown, 1998), and as a means of expanding dictionary coverage (Baldwin et al., 1999). There has also been some work on automatically extracting verbs which undergo particular alternations from corpora (McCarthy and Korhonen, 1998). Our work represents a variation on this same theme, whereby we compare the different *senses* of a given verb within the Goi-Taikei pattern-based valency dictionary (Ikehara et al., 1997; Shirai et al., 1997), and exhaustively determine all possible mappings between each valency frame pair. Naturally, not all such mappings are going to constitute true alternations, but by scoring each mapping and combining the scores for each mapping paradigm over all dictionary entries, it is hoped that alternations will find themselves into the top-ranking analyses. One area in which this research attempts to break new ground is, therefore, the fully-unsupervised extraction of alternations.

Our purpose in extracting alternations is twofold. Firstly, we are interested in determining the possibilities for collapsing dictionary entries together through dynamic generation of regular alternations. Secondly, we are interested in deriving the inventory of Japanese alternations through automatic means, for comparison against previous analyses of alternations in Japanese.

We test out three main methods to extract alternations, the first being based on full coincidence of selectional restrictions (method 1 – $\langle m_1 \rangle$), the second incorporating semantic backing-off to capture slight disagreements in selectional restrictions (method 2 – $\langle m_2 \rangle$), and the third scoring each alternation according to the quality of the match, in terms of the strength of the selectional restrictions and degree of overlap (method 3 – $\langle m_3 \rangle$). For each, we test the effects of exhaustive vs. highest-ranking analysis between case frames, and apply Zipf's law in weighting each case frame for expected frequency.

In the next section (Section 2), we define what we mean by alternation and outline the assumptions underlying this research. We then go on to describe the basic extraction methodology in Section 3, and detail the three proposed methods in Sections 4 to 6, describing the results obtained for each approach as we go. Section 7 concludes the paper with a discussion of the future direction of this research.

[1] Throughout this paper, we will tend to talk exclusively of selectional restrictions, where we mean both selectional restrictions and lexical fillers. Coincidence of selectional restrictions, therefore, refers to full coincidence of the range of lexical fillers as well as agreement in the selectional restriction content. Backing-off, on the other hand, applies only to selectional restrictions.

2 Definitions & assumptions

We define **alternation** as:

> *systematic inter-case frame variation operating at the* ***case marker, (canonical) word order*** *and/or* ***valency*** *levels, and also possibly involving adverbials and reflexivisation*

Alternation is thus a directed binary relation, linking non-coincident case frames. We term the pair of case frames undergoing alternation an **alternating pair**, from the **base case frame** to the **target case frame**. This process can be illustrated by the following examples:[2]

(1) a. *A-ga* *B-o* *hazimeta*
A-NOM B-ACC started$_{\text{TRANS}}$
"A started B"

b. *B-ga* *hazimatta*
B-NOM started$_{\text{INTRANS}}$
"B started"

(2) a. *A-ga* *B-o* *kaisi-sita*
A-NOM B-ACC started$_{\text{TRANS}}$
"A started B"

b. *B-ga* *kaisi-sita*
B-NOM started$_{\text{INTRANS}}$
"B started"

(3) a. *A-ga* *naita*
A-NOM cried
"A cried"

b. *B-ga* *A-o* *nakaseta*
B-NOM A-ACC made cry
"B made A cry"

(4) a. *A-ni* *B-ga* *dekiru*
A-DAT B-NOM can do
"A can do B"

b. *A-ga* *B-o* *dekiru*
A-NOM B-ACC can do
"A can do B"

Example (1) comprises the lexical causative alternation, (2) the (analytic) causative/inchoative alternation and (3) the synthetic causative alternation, each taking the direction base case frame (a) to target case frame (b). Note that the first two of these are analysed as valency-reducing alternations, but the last as a valency-increasing alternation. Our reason in this is the general applicability of the synthetic causative and markedness of synthetic causative verb inflection. Additionally, the first two alternation types operate over a closed set of verbs, with localised selectional preferences for alternating case slots, whereas the synthetic causative is applicable for a much less well-defined set of verbs and introduces an argument of essentially fixed selectional restrictions. Predictability of affected argument type and the ability to describe an alternation by way of a generalised template are requirements on alternations.

Example (4) provides an example of a valency-preserving case marker alternation.

We make the assumption that alternations producing modification of the argument structure will either

[2] The following abbreviations are used in sentence glosses: NOM = nominative, ACC = accusative, DAT = dative, # = semantically infelicitous.

add or delete arguments, with these two processes never occurring simultaneously within a single alternation. That is we can never have, for example, a valency-preserving, argument structure-modifying alternation. We additionally stipulate that the case frame must be modified in some way under alternation.[3]

By associating a direction with each alternation, we are forced into stipulating which of the two affected case frames is basic, a question which is clearly beyond the capability of an automatic extraction system. For the time being, we get around this by enforcing the constraint that alternations must be either valency-preserving (i.e. not modify the argument structure) or valency-reducing. Additionally, for valency-preserving alternations, we normalise the direction of the alternation so as to be able to combine analysis of all like alternations together. In effect, all this achieves is to remove representational multiplicity and allow us to maximally cluster like alternations together in determining the overall permeance of each alternation type, as all valency-increasing alternations are captured in reverse form as valency-decreasing alternations. Admittedly, however, this does produce misrepresentation in that the synthetic causative from (3) above, for example, is analysed as occurring in reverse to the direction claimed above.

This research relies heavily on the following **Assumption of Preservation of Selectional Restrictions** ("APSR"), originally proposed in Baldwin et al. (1999):

> *Selectional restrictions are essentially unchanged under alternation, with any idiosyncratic sub-preferences being defeasible given appropriate context*

In this, we do not seek to refute the "Principle of No Synonymy of Grammatical Forms" (Goldberg, 1995, p 3) that different syntactic realisations necessarily produce different meanings, that is that the members of an alternating pair display some difference in meaning or focus/topicality. Rather, we make the claim that corresponding case slots in the two case frames will display the same basic range of case fillers.

It is relatively easy to fashion what would appear to be counter-examples to this claim.

(5) *I rolled the box up the slope*

(6) #*The box rolled up the slope*

(7) *The anti-gravitational ball rolled up the slope*

(8) *The box came hurtling down the hill and rolled up the slope*

Despite the high acceptability of the transitive alternate of *roll* in (5), infelicity is produced for the unaccusative alternate in (6). However, closer observation of these sentences reveals that the issue at stake here is not whether the subject of the unaccusative *roll* can be instantiated with *box*, but that the act of rolling must be inherently facilitated. Hence, given a subject such as an *anti-gravitational ball* which has the innate ability to move up slopes, (7) becomes perfectly acceptable, and similarly, given a context in which we can see that the box has sufficient momentum to roll up a slope, felicity is produced as seen in (8). As such, the factor blocking the acceptability of (6) is not so much selectional restrictions as pragmatic interpretation.

[3] In extraction, this final stipulation filters off intra-case frame alternations of the form $[(c_\alpha \rightarrow c_\beta)\ (c_\beta \rightarrow c_\alpha)]$, where the selectional restrictions on case slots c_α and c_β are identical.

A more serious threat to the validity of the APSR is the effects of definiteness and volitionality, and the observation that alternations which focus indefinite arguments over definite arguments, or involitional arguments over volitional arguments often produce ungrammaticality. Also, overly long arguments (in terms of linear length) sometimes alternate where shorter arguments with the same basic semantic head do not alternate, or vice versa. Here, we take the line that universal grammatical and lexical principles can overrule basic alternations, but that given the correct definiteness value, volitionality and linear length of arguments, alternating case slots will display the same range of filler types in all realisations.

3 Basic methodology

The basic methodology employed to extract alternation candidates is to take a verb dictionary annotated with selectional restrictions, exhaustively determine correspondences between lexically related items, and look for regularities in the patterns of correspondence. By 'lexically related items' is meant that the two verbs in question must share a kanji prefix. This greatly reduces the search space and at the same time maintains a tenuous semantic link between items considered for alternation.

3.1 The dictionary

The particular dictionary targeted for extraction is the verb component of the Goi-Taikei pattern-based valency dictionary, comprising 5241 verb tokens and 13822 case frame instances. Each case frame contains one or more case slots, each annotated with lexical fillers and/or selectional restrictions, with the case slots listed in canonical order. One aspect of the Goi-Taikei valency dictionary which we rely on to weight case frames according to expected frequency of occurrence, is the listing of the case frames in order of sense saliency. That is, the first-ranking entry for a given verb type is the most prominent sense for that verb, the second-ranking entry the second most prominent sense, and so on.

The selectional restriction annotation is linked to the Goi-Taikei thesaurus by way of one or more nodes in the thesaurus tree structure.

As described above, this whole stream of research is founded on the assumption that alternations are characterised by preservation of selectional restrictions. In terms of the extraction process, we must have implicit faith not only in our hypothesis but also the consistency of the lexicographers who created the dictionary. We return to this point in the discussion of the various extraction techniques.

Another precondition on the success of this research is that the target valency dictionary contains a full spectrum of alternations for us to unearth. We expect more prominent alternations to get better coverage partly because of their genuine commonality, but also because of their salience in the minds of a lexicographer when adding an entry undergoing that alternation to the valency dictionary. We should therefore not be surprised to see very high frequency for core alternations, and much lesser frequency for relatively peripheral alternations.

The Goi-Taikei pattern-based valency dictionary was originally designed for use with the ALT-J/E transfer-based machine translation system (Ikehara et al., 1991). This has implications for this research in terms of the criteria used to partition off case frames (i.e. verb senses), in that polysemous verbs with a correspondingly polysemous English translation will generally be described as a single case frame. In the event that the alternates of the Japanese case frame are associated with distinct English translations, multiple alternates will be produced, with selectional restrictions tuned to differentiate usage of the alternates. In cases such as

this, therefore, selectional preferences will not be preserved under alternation, due to inter-lingual idiosyncrasies rather than a genuine clash with our APSR.

Perhaps a more immediate concern is what to do in cases of analytical ambiguity, such as when multiple case slots in one case frame agree with the selectional restrictions on one or more case slots in the other case frame. Here, how do we determine which of the analytical possibilities is correct? Below, we consider different methods for resolving such ambiguity.

3.2 Representing alternations

Extracted alternations are represented as the tuple $\langle [SUFF], [ARGS], SCORE \rangle$, where *SUFF* is the directed non-corresponding suffix component between alternates, *ARGS* is the directed case slot correspondence and *SCORE* is a numeric statement of the plausibility of the alternation, $SCORE \in [0,1]$. For example, we represent the alternation (2a) → (2b) from above as:[4]

$\langle [\text{-}], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 1 \rangle$

That is, the verbs for the base and target case frames coincide, the base nominative (*ga*) case slot is deleted, and the base accusative (*o*) case slot alternates with the target nominative (*ga*) case slot. Note that *ARGS* presents the case slots in the order given in the base case frame, and the case marker mapping of each onto the target case frame.

4 The basic model – $\langle m_1 \rangle$

Method 1 ($\langle m_1 \rangle$) is based around full match of selectional restrictions for case frames with a common kanji stem.

4.1 The $\langle m_1 \rangle$ extraction procedure

For each pair of case frames where the respective verbs share some common kanji prefix, we exhaustively align case slots which are identical in selectional restrictions and lexical fillers, and at the same time allow for the possibility of case slots having been deleted from one of the case frames. All mappings which produce a single match for every case slot contained in one of the two case frames, are returned as alternations from the second case frame to the first (hence maintaining our constraint on valency-increasing alternations).

We are immediately faced with the issue of analytical ambiguity and what to do in cases of multiple potential alternations existing between case frames. We examine two techniques to deal with this situation: (1) ignore case frame pairings which involve analytical ambiguity and allocate a *SCORE* of 1 for each remaining alternation (**ambiguity ducking**); and (2) allocate a *SCORE* of $\frac{1}{n}$ for each potential alternation between a given case frame pairing, in the case of *n*-fold ambiguity (**score normalisation**).

The following is example output for the case of score normalisation:

$\langle [\text{-}], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 1 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\}\rightarrow\phi)], 0.5 \rangle$
$\langle [\text{-}], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 0.5 \rangle$

The first line represents an outright causative/inchoative alternation, whereas the second and third lines represent the instance of ambiguity between the causative/inchoative and unspecified object alternation types (hence the respective scores of 0.5 as compared to 1 for the first line), caused through the two case slots in the base case frame having the same selectional restrictions.

One other area in which the described formulation is lacking, is that it has no way of determining the frequency of each case frame, and hence the relative conspicuity of any alternations between them. That is, we want to be able to say that alternating case frames in common use will produce greater salience for that alternation than alternations between rarely-occurring case frames. To determine the true frequency of each case frame we would have to employ some verb sense disambiguation method, inevitably introducing more noise into the task. Instead, therefore, we focus on the expected relative frequency of occurrence of each case frame for a given verb, and rely on the ordering of case frames within the original valency dictionary to weight each case frame accordingly. For this purpose, we apply **Zipf's law** which states that the frequency f of a word (or in our case, case frame) is directly proportional to the relative rank r of that word, or in other words, that there is some constant k such that $f \cdot r = k$. In application terms, therefore, we allocate the top-ranking case frame a weight zw of 1, and weight each subsequent case frame by a factor of $\frac{1}{r}$, where r is the rank of that case frame. The *SCORE* for an alternation from case frame a to case frame b, is then weighted by $zw_a \cdot zw_b$.

4.2 Alternation clustering

Having produced a listing of all legal alternations between case frames, we next cluster them together to determine the distribution of each alternation type.

The first step in this process (**step 1**) is to combine together identical alternations sharing the same *SUFF* component, and add together the respective scores. This simple process is sufficient to cluster together analytic, but not synthetic and lexical alternations.

The next step (**step 2**) is to collapse together all combined alternations from step 1, for which the *SUFF* component is covered by a single lexical/conjugational paradigm. Lexical paradigms are a classification of transitive/intransitive and causative/non-causative verb pairs according to derivational affix. An example of such a derivational affix pair is *-e-/-ar-*, as seen for such verbs as *ageru/agaru* "to raise"/glrise and *sonaeru/sonowaru* "to provide"/"be endowed with" (see Jacobsen (1992, pp 258–68) for a full listing of such affix pairs). The only conjugational paradigms currently considered are the passive and synthetic causative. All alternations governed by a common lexical/conjugational paradigm are clustered together into one common alternation, with *SUFF* describing the paradigm applied in the clustering process. Note that there is no overlap between the particular paradigms currently targeted, such that ambiguity as to the applicable paradigm type can never occur.

In the final step of alternation clustering (**step 3**), we score up "sub-alternations" based on the output of step 2, by identifying the core argument content of individual alternation instances/clusters. This entails iteratively stripping off the final case slot from the end of *ARGS* in the alternation $\langle [SUFF], [ARGS], SCORE1 \rangle$, and, in the case that the deleted case slot cs_i is non-alternating and the derived alternation $\langle [SUFF], [ARGS \ominus cs_i], SCORE2 \rangle$ has been observed in the data, incrementing *SCORE2* by *SCORE1*. This procedure is halted once we reach a non-alternating final case slot. That we choose the final case slot for deletion is based on the assumption that this is the most peripheral argument contained in the argument list, given that the argument list is described in canonical order. The stripping of case slots is a non-destructive process, such that the original alternation is preserved in the final data but the scores

[4] A *SUFF* value of '-' indicates that the verb lexicalisation is unchanged under alternation (i.e. the alternation is analytic).

of sub-alternations are bumped up. An example of the above process can be seen in the removal of case slot 3 from the following case frame, to produce the causative/inchoative sub-alternation.

$$[(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})\ (NP_3\{kara\})] \\ \Longrightarrow [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})]$$

4.3 Results for $\langle m_1 \rangle$

We were able to generate 2831 alternation candidates using the above method, which clustered into 1520 alternation tokens. Note that the total number of case frame pairings was 590,702, such that we produced an alternation for about five in every thousand case frame pairings on average.

Looking first to the output of the general method, with score normalisation but without applying Zipf's law, we find the causative/inchoative and unspecified object alternations to be by far the most common alternation types, with around 133 occurrences each. These are followed by quotative alternations, with sentential complements alternating between the accusative/dative and quotative cases (corresponding to indirect and direct speech, respectively):

$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\}\rightarrow\phi)], 133.7 \rangle$
$\langle [\text{-}], [(NP_1\{ga\}\rightarrow\phi)(NP_2\{o\}\rightarrow\{ga\})], 133.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (S_2\{to\}\rightarrow\{ni\})], 66.5 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (S_2\{to\}\rightarrow\{o\})], 62.7 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{ni\}\rightarrow\phi)], 38.2 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{kara, yori\}\rightarrow\{o\})], 30.2 \rangle$
$\langle [e/ar], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 28.5 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{ni\}\rightarrow\{o\})], 27.2 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\})\ (NP_3\{to\}\rightarrow\{ni\})], 19 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{kara, yori\}\rightarrow\phi)], 15.7 \rangle$

Surprisingly few lexical or synthetic alternations were found in the top-ranking alternations, with only 7 of the top 30 alternations being non-analytic. The translingual nature of the valency dictionary came out in the fact that the passive alternation was ranked 12th and the synthetic causative alternation ranked 19th (i.e. we would not expect synthetic alternations to occur as separate entries in a mono-lingual dictionary unless there was some departure in meaning/usage between the alternates).

The scarcity of alternations in data is highlighted by the alternation at rank 30 having a frequency of 8 with score normalisation activated, especially given the total number for possible verb frame combinations. We are down below a frequency of 5 beyond rank 51, and even here, many of the higher ranking alternations are either dubious or super-alternations of other higher-ranking alternations (i.e. alternations which have contributed to the increased count of alternations above them through clustering).

The relative frequency of the unspecified object alternation seems incredibly high, suggesting that there had been analytical ambiguity between the causative/inchoative unspecified object alternations at various points. Inspection of the data with ambiguity ducking bears out this prediction to some degree, with the causative/inchoative alternation outscoring the unspecified object alternation 107 to 90, but these two alternation types still occupying the top two positions. Otherwise, the relative ordering of alternations with ambiguity ducking is almost identical to that with score normalisation.

Evoking Zipf's law in weighting case frame pairings produces an almost identical ranking of alternations to the basic method, although with the causative/inchoative coming out on top as for the ambiguity ducking method. Further analysing the mean score for each alternation type of frequency 5 or more, we get rather different data, as detailed below:

$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{ni, e\}\rightarrow\phi)], 0.4 \rangle$
$\langle [\text{-}], [(NP_1\{ga\}\rightarrow\phi)(NP_2\{o\}\rightarrow\{ga\})(NP_3\{ni\})], 0.4 \rangle$
$\langle [as/e], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 0.4 \rangle$
$\langle [e/ar], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 0.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\}\rightarrow\phi)\ (NP_2\{o\}\rightarrow\{ga\})], 0.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{ni\}\rightarrow\{o\})], 0.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{to\}\rightarrow\{o\})], 0.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{kara, yori\}\rightarrow\{o\})], 0.3 \rangle$
$\langle [as/u], [(NP_1\{ga\})\ (NP_2\{o\}\rightarrow\phi)], 0.3 \rangle$
$\langle [e/ar], [(NP_1\{ga\}\rightarrow\phi)(NP_2\{o\}\rightarrow\{ga\})(NP_3\{ni\})], 0.3 \rangle$

As can be seen from these top 10 alternations, we get a higher number of lexical alternations (8 in the top 20), and of the type expected. For analytic alternations also, we are getting more of the sorts of alternations we had expected. All of the the alternations have reasonable applicability as well, with the average frequency being 40.7.

The unexpected sparseness of alternation analyses produced by $\langle m_1 \rangle$ may have been due to its inability to pick up on alternations which diverged only slightly in selectional restrictions due to annotational inconsistency in the dictionary. That is, it may be the case that what should have been recognised as alternations and hence have matching selectional constraints were overlooked, a possibility we consider by way of $\langle m_2 \rangle$.

5 Backing-off – method 2

In method 2 ($\langle m_2 \rangle$), we seek to consolidate on the results achieved for $\langle m_1 \rangle$ by covering the same basic set of alternation instances, but also including extra alternation instances through semantic backing-off. $\langle m_2 \rangle$ is identical to $\langle m_1 \rangle$ in all respects other than $\langle m_2 \rangle$ allowing for backing-off, and adjusting the *SCORE* of the affected alternation according to the degree of backing-off.

5.1 The $\langle m_2 \rangle$ extraction procedure

Semantic backing-off is a method of relaxing selectional restrictions, and in this case refers to the minimal relaxation of the selectional restrictions in the base and target case slots required to produce coincidence. As described above, selectional restrictions are described as a set of nodes in the tree structure of the Goi-Taikei thesaurus. Given two case slots c and d, we match up the selectional restrictions by taking each node c_i from the selectional restriction description of c and attempting to match it in the selectional restriction description of d. In the case of a perfect match not being attained, we search for the closest node to c_i. This is achieved by determining the **lowest common supernode** ("LCS") of c_i and each d_j described in the selectional restrictions for d, and returning a score describing the relative effort required to coerce c_i and d_j to that LCS.

We define the penalty $pen(s_1, s_2)$ in matching selectional restrictions s_1 and s_2 together as in equation (1), where $super(s_1, s_2)$ returns the lowest common supernode of s_1 and s_2, and $depth(n)$ returns the depth to node n as the number of nodes in the path from the thesaurus root to n, inclusive. In this, we penalise according to the combined number of levels of the tree s_1 and s_2 have been cranked up through, with a penalty of

$$pen(s_1, s_2) = 1 - \frac{depth(super(s_1, s_2))}{depth(s_1) + depth(s_2) - depth(super(s_1, s_2))} \quad (1)$$

$$comb\_pen(a, b) = \sum_i \min pen(a_i, b_j) + \sum_{b_j \notin A_{min}} \min pen(a_i, b_j) \quad (2)$$

0 if no backing-off is required, and $pen(s_1, s_2)$ tending towards 1 as the degree of relaxation increases. Note that we punish relaxation into the upper reaches of the tree over that at the level of lower branches, as the upper levels are expected to desecrate the original selectional restrictions to a greater degree than would occur further down the tree structure.

The combined penalty for case slots a and b with selectional restrictions $a_1, a_2, ...a_k$ and $b_1, b_2, ...b_l$ is calculated as in equation (2), where A_{min} is the set of nodes b_m producing a minimal penalising edge from each a_i. In this way, we enforce the requirement that all nodes described in the selectional restrictions for both a and b must match with some node in the opposing case slot, but at the same time alleviate duplication of penalty.

We are now in a position to determine the overall penalty across all corresponding case slots for a given alternation. The overall score for case slot alignment $(a_{\theta_1} - b_{\kappa_1})$ $(a_{\theta_2} - b_{\kappa_2})$ ... $(a_{\theta_m} - b_{\kappa_m})$ between case frames A and B is given by:

$$\frac{1}{(1 + \sum_{i=1}^{m} comb\_pen(a_{\theta_i}, b_{\kappa_i}))^{\alpha}} \quad (3)$$

where α is a weighting constant, $\alpha > 0$. We provisionally set α to 2 in evaluation.

As for $\langle m_1 \rangle$, we investigated various combinations of ambiguity reduction techniques and weighting according to Zipf's law. In this case, ambiguity ducking was carried out by returning only the absolute top-ranking alternation analysis for the current case frame pairing, and not not committing ourselves to any analysis in the case of multiple top-ranking alternation candidates.

The alternation candidates were clustered together as for $\langle m_1 \rangle$ (see Section 4.2).

5.2 Results for $\langle m_2 \rangle$

$\langle m_2 \rangle$ produced a total of 163,282 alternation candidates, describing a total of 49,685 alternation tokens. As for $\langle m_1 \rangle$, the causative/inchoative and unspecified object alternations well out-scored other alternation types, but with other analysis types following along close behind.

$\langle [\text{-}], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})], 253.1 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)], 224.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{ni\} \rightarrow \phi)], 143.4 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\})\ (NP_3\{ni\} \rightarrow \phi)], 119.1 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)\ (NP_3\{ni\} \rightarrow \{o\})], 115.7 \rangle$
$\langle [\text{-}], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{ni\} \rightarrow \{ga\})], 115.4 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{to\} \rightarrow \{o\})], 115.0 \rangle$
$\langle [e/ar], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})], 105.6 \rangle$
$\langle [e/ar], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)], 96.1 \rangle$
$\langle [\text{-}], [(NP_1\{ga\} \rightarrow \{o\})\ (NP_2\{o\} \rightarrow \phi)\ (NP_3\{ni\} \rightarrow \{ga\})], 83.6 \rangle$

Overall, lexical alternations seemed to be pushed up the ranking slightly and synthetic alternations down, with analytic alternations bumped up or down almost randomly. This supports our initial hypothesis that prominent analytic alternations will get artificially high coverage. It also points to a high degree of inconsistency of selectional restrictions between lexical alternates, as backing-off picked up a disproportionate number of alternations fitting into the recognised lexical paradigms.

The relatively high ranking of the improbable unspecified object *-e-/-ar-* alternation is an artefact of over-allowance of backing-off, as could be reduced by lowering the value of α.

Running $\langle m_2 \rangle$ with weighting according to Zipf's law, produced the expected result of scoring up lexical alternations, although not to the degree expected.

On re-ranking the output alternations according to the quality of match rather than cumulative score (for alternations occurring at least five times), an rather different composition of alternations was produced:

$\langle [suru/rieru], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})], 0.3 \rangle$
$\langle [e/ar], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})\ (NP_3\{ni, e\})], 0.3 \rangle$
$\langle [me/maseru], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)], 0.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})\ (NP_3\{kara, yori\})\ (NP_4\{ni, e, made\})], 0.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)\ (NP_3\{to\})], 0.3 \rangle$

We seem to get a high number of low-frequency novel synthetic and lexical alternations, with the core lexical alternations scoring badly due to the high numbers of poorly matching alternations. This again reinforces our claim as to annotational inconsistency of selectional restrictions between lexical alternations

6 Semantic density – method 3

In method 3 ($\langle m_3 \rangle$), we take a slightly different tack to $\langle m_1 \rangle$ and $\langle m_2 \rangle$, in evaluating not only the degree of difference between selectional restrictions, but also the quality of the match. This is achieved through the notion of "semantic density".

6.1 The $\langle m_3 \rangle$ extraction procedure

As for $\langle m_1 \rangle$ and $\langle m_2 \rangle$, we consider only case frames which share a common kanji stem, and exhaustively generate all possible alternations between them. We also allow for semantic backing-off in a manner similar to $\langle m_2 \rangle$. With $\langle m_3 \rangle$, however, we balance up the degree of backing-off against the quality of the produced match, such that matches in the lower reaches of the thesaurus structure are scored higher than those matching higher up. This is achieved by scoring matches according to the semantic density of the region described by the matched selectional restrictions.

Semantic density is an indication of the degree of specificity of a given semantic region, as determined according to the thesaurus topology. Consider Fig. 1 below. Here, region R_3 is more dense than regions R_2 or R_1, in the sense that concepts contained within R_3 would have greater cohesion that those in the other two regions. If selectional restrictions are given this degree of specificity, it must mean that the associated case slot is highly specialised in its usage and that the lexicographer encoding the selectional restrictions is confident as to the demarkation of use of that case slot. A match at this high level of specialisation tends to have greater credibility than a match at a higher level, and point to genuine case slot correspondence.

We model semantic density-based match quality according to **case slot restrictiveness** (CSR) (Baldwin and Tanaka, 1999). The degree of CSR of a given node

$$comb\_score(a,b) = \sum \min_i pen(a_i,b_j)^\alpha \times CSR(super(a_i,b_j)) + \sum \min_{b_j \notin A_{min}} pen(a_i,b_j)^\alpha \times CSR(super(a_i,b_j)) \quad (4)$$

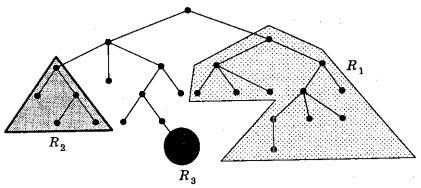

Figure 1: Semantic density

x subsuming nodes $l_1, l_2, ...l_n$, is estimated as:

$$CSR(x) = \frac{n}{\sum_{i=1}^{n} tree\_depth(x, l_i)} \quad (5)$$

where *tree_depth* is defined as the number of nodes between the subtree root x and subsumed leaf l_i, inclusive. This produces the desired ranking for the above figure of $0 < CSR(R_1) < CSR(R_2) < CSR(R_3) \leq 1$.

We balance up the degree of semantic density against the degree of backing-off required to achieve that semantic density, using the *pen* function as for $\langle m_2 \rangle$. The combined score for case slots a and b with selectional restrictions $a_1, a_2, ...a_k$ and $b_1, b_2, ...b_l$ is calculated as in equation (4), where A_{min} is the set of in equation (2), where A_{min} is the set of nodes b_m producing a minimal penalising edge from each a_i, and α is a weighting constant, $\alpha > 0$. We provisionally set α to 3 in evaluation.

We calculate the overall score for alignment $(a_{\theta_1} - b_{\kappa_1})\ (a_{\theta_2} - b_{\kappa_2})\ ...\ (a_{\theta_m} - b_{\kappa_m})$ between case frames A and B simply as:

$$\sum_{i=1}^{m} comb\_score(a_{\theta_i}, b_{\kappa_i}) \quad (6)$$

As with $\langle m_2 \rangle$, we trialled removing ambiguity by taking the unique top-ranking analysis, and also weighting according to expected frequency of occurrence through multiplication by the zw weights described for $\langle m_1 \rangle$.

Clustering of the output data is carried out in an identical fashion to $\langle m_1 \rangle$ and $\langle m_2 \rangle$.

6.2 Results for $\langle m_3 \rangle$

$\langle m_3 \rangle$ produced a total of 155,906 alternation candidates, comprising 46,476 alternation tokens. Here again, the causative/inchoative and unspecified object alternations emerge as the highest-ranking alternation types in the cumulative score ranking, again due to their superior frequency of occurrence.

$\langle [\text{-}], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})], 205.8 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\})\ (NP_3\{ni\} \rightarrow \phi)], 151.8 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)], 142.6 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)\ (NP_3\{ni\} \rightarrow \{o\})], 142.0 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{to\} \rightarrow \{o\})], 111.7 \rangle$
$\langle [e/ar], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})], 98.3 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\})\ (NP_3\{ni\} \rightarrow \{ni, e\})], 77.4 \rangle$
$\langle [\text{-}], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\})\ (NP_3\{ni\} \rightarrow \{ga\})], 75.3 \rangle$
$\langle [e/ar], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)], 65.9 \rangle$
$\langle [\text{-}], [(NP_1\{ga\})\ (NP_2\{o\})\ (NP_3\{ni, e\} \rightarrow \phi)], 65.7 \rangle$

We see that $\langle m_3 \rangle$ has had trouble in differentiating certain analyses, such as the second alternation against the eighth ranking alternation. It should be possible to resolve such ambiguity to some degree by increasing the value of α. Lexical alternations appear to fare better with $\langle m_3 \rangle$ than the other two methods.

The inclusion of Zipf's law further helps out the case of lexical alternations, with the lexical causative ranking a creditable fourth, but the misleadingly high-scoring lexical unspecified object being scored down relatively.

We find an interesting effect on analysing the mean score for each alternation type (without Zipf's law), in that the five top-ranking alternations are all subtle variations on each other. Essentially what we have here is a causative/inchoative alternation, with unmodified local allative (case *ni/e/made*) and local ablative (case *kara/yori*) case slots (the optionality of the *made* marking in the unaccusative is seen as an error in annotation). Because of identical selectional restrictions on the two local case slots in each case frame, the system is unable to choose the correct combination thereof, and is additionally having to cope with the selectional restrictions on the subject and object case slots of the transitive case slot nearly coinciding. The correct ordering is produced in the end, but a higher α value would certainly produce greater discriminatory power for the system.

$\langle [gesuru/garisuru], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})\ (NP_3\{kara, yori\})\ (NP_4\{ni, e, made\})], 1.6 \rangle$
$\langle [gesuru/garisuru], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})\ (NP_3\{kara, yori\})\ (NP_4\{ni, e, made\} \rightarrow \{ni, e\})], 1.5 \rangle$
$\langle [gesuru/garisuru], [(NP_1\{ga\} \rightarrow \phi)\ (NP_2\{o\} \rightarrow \{ga\})\ (NP_3\{kara, yori\} \rightarrow \{ni, e\})\ (NP_4\{ni, e, made\} \rightarrow \{kara, yori\})], 1.5 \rangle$
$\langle [gesuru/garisuru], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)\ (NP_3\{kara, yori\})\ (NP_4\{ni, e, made\})], 1.4 \rangle$
$\langle [gesuru/garisuru], [(NP_1\{ga\})\ (NP_2\{o\} \rightarrow \phi)\ (NP_3\{kara, yori\})\ (NP_4\{ni, e, made\} \rightarrow \{ni, e\})], 1.4 \rangle$

7 Discussion

It is difficult to draw any hard conclusions from the overall results as we have no way of empirically evaluating the quality or types of alternations produced. Intuitively speaking, we would have expected lexical alternations to play a more prominent role. As mentioned above, one reason for their low profile is annotational inconsistency, as evidenced by the different rankings of lexical alternations with and without semantic backing-off. One other factor alluded to above, is that the dictionary was always intended for machine translation, making translational sufficiency the absolute evaluation criterion on dictionary quality. Thus, that *hazimeru* should have a single entry with English translation "start$_{TRANS}$", but its lexical unaccusative alternate *hazimaru* have fully 21 separate entries each with a distinct English translation,[5] is perhaps not disturbing to

[5] Admittedly, 17 of these were targeted at the stockmarket domain, but there were still four generalised entries to *hazimaru*'s

the system developer. It certainly does not help our case, however.

It is certainly true that the different methods were able to produce different effects, particularly in the ranking of average scores for each alternation type. The combined scores produced by $\langle m_1 \rangle$ are perhaps most indicative of the true coverage of the different alternations in the dictionary, the mean score for $\langle m_2 \rangle$ the most indicative of the shakiness of certain alternations, and the mean score for $\langle m_3 \rangle$ the most indicative of true alternation quality. One clear result to appear from all methods was that the causative/inchoative and unspecified object alternations are the two most pervasive alternation types. This is with the caveat that the high salience of the unspecified object alternation can be put down to, again, the leaning of the dictionary towards machine translation, and non-existence of an English verb taking that same alternation.

The observant reader may have been surprised at the vast number of alternation tokens produced for $\langle m_2 \rangle$ and $\langle m_3 \rangle$ in particular. The main contributing factors here are: (a) analytical ambiguity, and the system returning all possible mappings between case frames, irrespective of plausibility; and (b) the inclusion/exclusion of middle case slots producing multiple variants of the same basic case frames. Analytical ambiguity can be ignored for the large part, due to implausible alternations being heavily penalised. Alternation clustering removes the effects of middle case slots in the cumulative frequency ranking, although we found them appearing in mean score rankings.

Looking to the future, the next step in this research will be to apply the extracted alternations in alleviating dictionary entries reproducible from the base case frame and an alternation description. We are also interested in looking at the patterns of selectional restriction variation under alternation, to determine whether the inconsistency seen for lexical alternations, for example, is highly regular and goes against our APSR. Either way, we should be able to detect errors in selectional restrictions for use in tuning the dictionary.

A further area in which we hope to apply this research is in the identification and categorisation of both lexical and semantic correspondences in the data. For example, it should be possible to postulate alternating lexical pairs through the degree of correspondence of their various case frames. To take this process one step further, it should also be possible to identify alternating semantic pairs such as *buy/sell* and *win/lose*, by relaxing our constraint on alternates having to share a common lexical stem.

Acknowledgements

The authors acknowledge the crucial role the Goi-Taikei resources played in this research, and express their gratitude towards the NTT translation communication research group for providing access to them. On a personal level, vital input was received from Christoph Neumann (TITech), Francis Bond (NTT) and participants of the 3rd Morphology/Lexicon Forum at Osaka University.

References

T. Baldwin and H. Tanaka. 1999. Argument status in Japanese verb sense disambiguation. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 207–217.

T. Baldwin, B. Hutchinson, and F. Bond. 1999. A valency dictionary architecture for machine translation. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 207–217.

B.J. Dorr and M.B. Olsen. 1996. Multingual generation: The role of telicity in lexical choice and syntactic realization. *Machine Translation*, 11:37–74.

N. Fukui, S. Miyagawa, and C. Tenny. 1985. *Verb Classes in English and Japanese: A Case Study in the Interaction of Syntax, Morphology and Semantics.* Lexicon Working Papers #3, Center for Cognitive Science, MIT.

A.E. Goldberg. 1995. *A Construction Approach to Argument Structure.* The University of Chicago Press.

K.L Hale and S.J. Keyser. 1987. *A View from the Middle.* Lexicon Working Papers #10, Center for Cognitive Science, MIT.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, pages 101–106, Washington DC. (http://xxx.lanl.gov/abs/cmp-lg/9510008).

S. Ikehara, M. Miyazaki, A. Yokoo, S. Shirai, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei – A Japanese Lexicon.* Iwanami Shoten. 5 volumes. (In Japanese).

W.M. Jacobsen. 1992. *The Transitive Structure of Events in Japanese.* Kurosio Publishers.

H. Jing and K. McKeown. 1998. Combining multiple, large-scale resources in a resuable lexicon for natural language generation. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics: COLING/ACL-98*, pages 607–13.

B. Levin. 1993. *English Verb Classes and Alterations.* University of Chicago Press.

D. McCarthy and A. Korhonen. 1998. Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98), Student Session*, pages 1493–5.

S. Shirai, S. Yokoo, H. Inoue, H. Nakaiwa, S. Ikehara, and A. Yagi. 1997. Nichi-ei kikai-hon'yaku ni-okeru imi-kaiseki no tame no kōbun jisho [A structural dictionary for semantic analysis in Japanese-English machine translation]. In *Proceedings of the Third Annual Meeting of the Japanese Association for Natural Language Processing*, pages 153–6. (In Japanese).

A. Wierzbicka. 1988. *The Semantics of Grammar.* John Benjamin.

one.