Alternation-based Lexicon Reconstruction

Timothy Baldwin* and Francis Bond†
* Center for the Study of Language and Information (CSLI)
210 Panama Street
Stanford, CA 94305-4115, USA
and
† NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0237, JAPAN
tbaldwin@csli.stanford.edu, bond@cslab.kecl.ntt.co.jp

Abstract

This research is aimed at developing a hierarchical alternation-based lexical architecture for machine translation. The proposed architecture makes extensive use of information sharing in describing valency frames through derivational links from base frames, rather than as independent entities. This has advantages in descriptive efficiency, robustness and maintainability.

The lexicon being developed is built up automatically from the Japanese component of an existing Japanese-English machine translation lexicon. The reconstruction process consists of analysing consistencies in selectional restrictions between valency frames, and postulating alternations where selectional restrictions are preserved on matching case slots; this was found to perform at 60.9% accuracy. All alternation candidates are incorporated into the final-version lexicon as derivational links, and expanded out at run time.

1 Introduction

This paper draws together and expands upon previous word on the reconstruction of a Japanese–English valency dictionary (Baldwin et al. 1999) and the automatic extraction of alternating case frames from a dictionary (Baldwin & Tanaka 2000). It aims to derive a hierarchical valency dictionary drawing heavily on explicit description of verbal alternation and with minimal descriptive redundancy, from a valency dictionary made up of simple transfer pairs. Due to heavy information sharing in the hierarchical valency dictionary, it benefits from enhanced maintainability, and can be updated and added to with much less input than a conventional transfer dictionary. At the same time, the hierarchical dictionary can be expanded out at run time, to capture the full potential and efficiency of the original dictionary.

One key objective of this research is to get away from a conventional transfer architecture in separating apart the source and target languages (Japanese and English, respectively) into independent lexicons. A "linking lexicon" is used to determine translational correspondence between the two lexicons, in the form of transfer links. This paper is focused particularly at the construction of the source language (i.e. Japanese) lexicon, although the design and means for restructuring the target language lexicon are essentially the same as described here. For details of the linking lexicon, the reader is referred to Baldwin et al. (1999).

Currently, the dictionary reconstruction process is fully automated and does not make use of bootstrap or any other external data, i.e., it is fully unsupervised. This

unsupervision extends to the set of alternation types featuring in our lexicon. That is, we do not start with a fixed set of attested alternations (e.g. of the type of Levin's English alternation inventory (Levin 1993), or Jacobsen's list of basic Japanese alternations (Jacobsen 1992)). Clearly, for the final version lexicon to be 100% linguistically accurate and technically reliable, some post-editing of data will have to take place. For the time being, however, we are interested in determining how far we can get without human intervention and without altering the basic informational content of the dictionary. As long as we ensure that the lexicon derivation process is lossless, the resultant lexicon can be used for MT purposes with confidence, as we will have neither lost nor gained anything over the original valency dictionary for run-time purposes.

This research is targeted at the Goi-Taikei Japanese–English valency dictionary (Ikehara et al. 1997), as used in the ALT-J/E machine translation (MT) system (Ikehara et al. 1991). The Goi-Taikei valency dictionary describes each Japanese verbal expression as a case frame headed by the verb in question. Each case slot is annotated with a discrete set of prototypical case markers, part of speech (NP or S), an obligatoriness flag, and a list of selectional restrictions and lexical fillers. The selectional restrictions take the form of nodes within the Goi-Taikei thesaurus tree. The Goi-Taikei thesaurus is an unbalanced tree of 2,710 nodes, connected by links showing either hyponymic (`is-a`) or meronymic (`has-a`) relations.

Given our primary interest in an enhanced source language lexicon, during the lexicon reconstruction process, we ignore the (English) translation associated with each transfer pair in the base dictionary, and consider only the source language component. Ultimately, the combined lexicons arising from this research are intended to interface seamlessly with ALT-J/E, providing the same informational content for MT purposes, but with greatly enhanced maintainability and consistency.

Alternations are crucial to the proposed dictionary structure. We define a (diathesis) alternation to be a 1-to-1 relation from a source to a target frame, which involves at least one of: (i) case marking variation between corresponding case slots, (ii) case slot deletion, and (iii) case slot insertion. To give an example, the causative-inchoative alternation occurs between a transitive and intransitive frame. In Japanese this involves both the deletion of the transitive subject case slot, and the mapping of the accusatively-marked (direct) object to the nominatively-marked subject position, as well as possible morphological transformation.[1] An example is shown in (1).[2]

(1) *Kim-ga* *doa-o* *aketa* / *doa-ga* *aita*
Kim-NOM door-ACC opened door-NOM opened
'Kim opened the door' 'The door opened'

This example illustrates the nature of case marking variation, the principal form of morphological variation considered in this research.

In its original form, the valency dictionary does not contain any explicitly-annotated alternations. Rather, alternants (e.g. *akeru* and *aku*) are described independently of one another, and any systematic variation that exists between them is incidental. The driving mechanism employed in the extraction of alternations is the assumption that the selectional restrictions associated with corresponding case slots are unchanged under alternation, originally proposed by Baldwin et al. (1999). That is, in the case of *akeru/aku*, the selectional restrictions associated with the direct object of *akeru*

[1]Note that for the purposes of this paper, we do not make explicit reference to the non-lexical effects of alternation, such as transformation of grammatical relation. We focus instead on the surface manifestation of such effects, e.g. in the form of case marking alternation.

[2]The following abbreviations are used in glosses: NOM = nominative and ACC = accusative.

"open$_{\text{trans}}$" are identical to those of the subject of *aku* "open$_{\text{intrans}}$", onto which it maps under alternation.

It is important to realise that, while we focus on verbal alternations throughout this paper, adjectival alternations also occur in Japanese (e.g. in the form of the double nominative construction) and feature in the final lexicon. We do not discuss them here simply because verbal alternation is more lexically and structurally diverse, and largely subsumes adjectival alternation.

The final structure of the dictionary is hierarchical, and draws on both inheritance and defaults. It is intended to minimise data duplication and maximise analytical sharing wherever possible. The resultant lexicon is thus greatly reduced in size over the base dictionary, but more importantly, any modifications to the lexicon tend to be localised, and additions can take advantage of structural parallelism and be highly abbreviated.

The remainder of this paper is structured as follows. Section 2 discusses theoretical issues and assumptions surrounding alternations. In Section 3, we then discuss the methodology employed to derive alternation data and briefly evaluate the alternation extraction method. In Section 4 we go on to describe the lexical representation, before concluding the paper with an overall discussion in Section 5.

2 Alternations: theoretical issues and assumptions

In the case of Japanese, alternation can be: (i) unmarked on the verb (**analytical** alternation, as seen for *hiraku* "open$_{\text{intrans/trans}}$"), (ii) marked on the verb stem by often-predictable lexical variation (**lexical** alternation, such as between *akeru* "open$_{\text{trans}}$" and *aku* "open$_{\text{intrans}}$"), (iii) marked by way of verbal inflection or a verb morpheme (**synthetic** alternation, such as occurs with the passive morpheme *(r)are*). Further, we are also investigating the utility of treating words with systematised alternation-type semantic correspondences but no morpho-syntactic marking (such as *kau* "buy" and *uru* "sell"), as alternations; this would add a fourth class: **cognitive** alternation. Fukui et al. (1985) treat certain combinations of verbs and affixes as alternations, we classify these as synthetic.

We make a number of assumptions about alternations in this research:

1. The selectional restrictions and lexical fillers on matching case slots are preserved under alternation
2. Alternations are monotonic in valency terms
3. A given alternation type has fixed direction

The first of these assumptions states that corresponding case slots in the two alternants of a given alternation token, display the same selectional restrictions and lexical fillers. That is not to say that the same distribution of lexicalisations will be observable for the two case slots (e.g. see Baldwin & Tanaka (2000) for details of the impact of pragmatics and facilitation on argument realisation), but rather that they have the same basic scope for instantiation.

The second assumption states that a given alternation type cannot involve both case slot insertion and deletion. That is, alternations must be strictly valency-reducing, valency-increasing or valency-preserving. A corollary of this assumption is that all case slots in at least one of the two alternant case frames must be linked to a case slot in the second alternant case frame (with all case slots in *both* case frames mapping to a case slot in the opposing case frame iff they are of equal valency). Our definition of

alternation (see above) implies that alternations are 1-to-1: a single case slot cannot be linked to more than one alternant case slot.

The third and final assumption constrains the direction of a given alternation type in all its realisations, and is intended to facilitate the unsupervised extraction and application of alternations rather than to be a universal truth about alternations. That is, we have no immediate means of determining for each alternation token which is the base and which the derived form. Our solution is to impose direction on the alternation type, and apply this to all instances thereof. This is achieved by stipulating that all alternations are either valency-decreasing or valency-maintaining, and arbitrarily normalising the direction of valency-maintaining alternations using the alphabetic order of the case markers on case slots which undergo modification.

Our constraint on alternation direction is clearly a misconstrual of the facts. Consider, for example, the causative-inchoative alternation in English. For *open*, it is relatively uncontroversial to say that the transitive realisation is basic, and the intransitive derived, as for something to open, some external force or agency is required. For *walk*, on the other hand, we would not like to say that the transitive (e.g. *Kim walked the dog*) forms the basis for the intransitive (e.g. *the dog walked*). The imposition of directionality is not intended as a predictor of whether the suggested derived form can exist in absence of the base form for a given verb, for example (see Dorr & Olsen (1996)). All it says is that given both realisations, we *a priori* define one to be the base and the other the derived form.

3 Alternation extraction method

Naturally, we need some way of getting at alternations in order to utilise them in the revised lexicon. This is achieved by first taking all pairs of case frames from the base valency dictionary, and identifying the most plausible (if any) alternation for each from the set of all possible valence-monotonic case slot mappings between them. We then analyse trends in the alternation data and feed this directly into the lexicon.

Alternation candidates are scored by evaluating the quality of match of selectional restrictions on corresponding case slots. Case slot match quality is rated empirically in the manner described below. We could also use the existence of identical English translations as an indicator that two candidates are related by way of alternation (cf. the valency frame generation procedure of Fujita & Bond (2002)). For example, *akeru* and *aku* are both linked to the English translation *open*.

3.1 Scoring case slot matches

The quality of match between case slots is quantitatively described by comparing the relative proximity of the selectional restrictions describing each, within the Goi-Taikei thesaurus tree. As stated above, selectional restrictions are provided as thesaurus node indices, and the greater the topological overlap between and conceptual cohesion within the regions described for the two case slots, the higher the match quality. This is intended to reflect the intuition that the higher the specificity of the selectional restrictions, the greater the confidence of the lexicographer in their integrity. Matches at higher levels of specificity are thus of higher quality than matches at lower levels of specificity, and conversely, *mis*matches at higher levels of generality are of less concern than mismatches at lower levels of generality (= higher levels of specificity).

The conceptual cohesion of the subtree described by a given node is modelled by way of the relative entropy of the token population of that region, as determined from corpus occurrence statistics (in the manner of Resnik (1999)). Nodes describing sparsely-populated subtrees are thus given higher weights than densely-populated subtrees, and

as we ascend the tree, the node weights decrease monotonically, right down to a weight of zero for the root node (n_o).

Lexeme token frequencies are calculated based on the EDR corpus (EDR 1995). In the case that a given lexeme found in the EDR corpus has more than one sense listing, the Goi-Taikei thesaurus provides a listing of sense salience. This is combined with Zipf's law to distribute the count over the sense inventory, such that for sense i of lexeme lex_p, the thesaurus node containing $lex_{p,i}$ is allocated a portion of $freq(lex_p)$ proportional to $\frac{1}{i}$. Having determined the sum token population of each node n_q, the conceptual cohesion $cohesion(n_q)$ is calculated by:

$$cohesion(n_q) = -\log P(n_q) = -\log \frac{\sum_{lex_{p,i} \in n_q} freq(lex_{p,i})}{\sum_{lex_{p,i} \in n_o} freq(lex_{p,i})} \quad (1)$$

Once we have determined the score for each node, we can evaluate the semantic proximity of nodes n_j and n_k as the relative disparity between the conceptual cohesion of each and their least common hypernym $sub(n_j, n_k)$:

$$classmatch(n_j, n_k) = 3\, cohesion(sub(n_j, n_k)) - cohesion(n_j) - cohesion(n_k) \quad (2)$$

In the case that n_j and n_k are coincident, $sub(n_j, n_k) = n_j = n_k$, such that the overall *classmatch* score becomes $cohesion(n_j) = cohesion(n_k)$. It is important to realise that *classmatch* can be negative in the face of high levels of backing-off up the tree structure in order to reach $sub(n_j, n_k)$.

Naturally, a single set of selectional restrictions can occur in the form of multiple thesaurus nodes. In matching a pair of selectional restrictions, we determine the spanning bi-partite mapping between them for which the mean *classmatch* score for connected sense nodes is maximised. The overall match between case slots S and T, with selectional restrictions $n_{S,1}, n_{S,2}, ...n_{S,m}$ and $n_{T,1}, n_{T,2}, ...n_{T,n}$, respectively, is defined to be:

$$slotmatch(S,T) = \max_{\tau \in \Upsilon} \frac{\sum_{connect_\tau(n_{S,j}, n_{T,k})} classmatch(n_{S,j}, n_{T,k})}{|\tau|} \quad (3)$$

where $\Upsilon = \{\tau : \forall n_{S,w} \exists n_{T,x}\ connect_\tau(n_{S,w}, n_{T,x}) \ \wedge \ \forall n_{T,z} \exists n_{S,y}\ connect_\tau(n_{S,y}, n_{T,z}\}$, and $connect_\tau(n_{S,j}, n_{T,k})$ denotes the fact that $n_{S,j}$ and $n_{T,k}$ are linked within spanning bi-partite selectional restriction mapping τ.

We are now in a position to determine the overall score for a given alternation. We base this score only on those case slots which are mapped from/to, i.e. deleted case slots do not enter into calculations. For case frames S and T, therefore, containing case slot mappings $(s_{\theta_1} - t_{\kappa_1})\ (s_{\theta_2} - t_{\kappa_2})\ ...\ (s_{\theta_m} - t_{\kappa_m})$, the score is:

$$score(S,T) = score\Big((s_{\theta_1} - t_{\kappa_1})\ (s_{\theta_2} - t_{\kappa_2})\ ...\ (s_{\theta_m} - t_{\kappa_m})\Big) = \sum_i slotmatch(s_{\theta_i}, t_{\kappa_i}) \quad (4)$$

Note that *score* is transitive, that is $score(S,T) = score(T,S)$. As noted above, the *classmatch* function can return a negative value, meaning that the overall score for a given alternation can be negative given sufficiently high discrepancy in selectional restrictions. If this occurs, that candidate alternation is disallowed.

A significant number of case slots incorporate lexical fillers either in parallel to, or to the exclusion of selectional restrictions. Lexical fillers are used either in fixed expressions such as *chie-o shiboru* "wrack one's brains (lit: wring knowledge)", or to make distinctions finer than those that can be made with the thesaurus. Both uses

describe highly specific lexical restrictions, which must largely coincide in membership for a given case slot matching to be considered valid. That is, for case slots S and T where at least one of S and T is annotated with lexical fillers, a case slot mapping is permitted iff the following inequality over lexical filler sets lex_S and lex_T, respectively, is satisfied:

$$\frac{|lex_S \cap lex_T|}{|lex_S \cup lex_T|} > \frac{1}{2} \tag{5}$$

Note that this evaluation of the overlap of lexical fillers does not affect the value of *score*, but is instead a hard constraint on the validity of a case slot mapping between S and T.

Despite our assumption that selectional restrictions on corresponding case slots are preserved under alternation (§ 2), we allow some degree of mismatch of both selectional restrictions and lexical fillers. Our motivation in this is to be able to cope with variability in the source lexicon. Due to the flat structure of the original valency dictionary, selectional restrictions and lexical fillers on corresponding case slots can vary considerably, without lexicographic motivation. A major aim of the lexicon reconstruction process is to identify any inconsistencies and then either resolve them or mark them as genuine differences.

One area where we both detect and resolve inconsistency is case marker alternation. Alternations where the case marker set of one case slot is a proper subset of that for the corresponding case slot, are treated as noise in the data rather than true artifacts of alternation, and the full case marker set used to supplement the impoverished set.[3]

3.2 Resolving alternation ambiguity

For each case frame pair, we return the best-scoring alternation(s), recalling that any negatively-scoring candidate alternations are automatically filtered off. In the case that a tie in score is produced, we select that candidate alternation which preserves case marking for the highest proportion of case slot mappings, under the assumption that alternations are conservative in their scope for modification. If the tie still remains, then we have no reasonable grounds for selecting between the candidate alternations, and no output is produced. With case frames of equal valency, it can happen that in the highest-scoring "alternation", no case marker variation has in fact taken place (i.e. the case frames are identical modulo linear reordering of case slots).[4] Based on our assumptions on the nature of alternations in Section 2, this does not construe a true alternation. Given that this is the best analysis for the given case frame pairing, we consider that no alternation exists.

3.3 Results of alternation extraction

A total of 2,777 alternation tokens were detected in the valency dictionary, from a total of 13,880 verbal case frames. Of these, 1,653 alternation tokens were incorporated into the revised lexicon (i.e. around 12% of verbal case frames are treated as being derived under alternation),[5] representing a total of 373 alternation types.

[3]There are a handful of instances where we retain the original case marking, such as {*ni*} matched with {*ni,kara*}, due to *ni* taking either a source (ablative) or target (allative) reading in isolation, but only a source reading when inter-replaceable with *kara*.

[4]In fact, the detection of identical case frames in the dictionary may prove valuable in the grander scheme of the lexicon overhaul. We would still not want to treat them as alternations, however.

[5]The disparity here is due to some case frames being mapped onto multiple times. In order to represent each case frame uniquely in the lexicon, 1,124 alternation tokens were pruned, by taking only the highest-scoring candidate alternation to each case frame.

Index	***Case slot mapping***			***Example verb***
1	$(NP_1\{ga\}\rightarrow\phi)$	$(NP_2\{o\}\rightarrow\{ga\})$		*kiru/kireru*
2	$(NP_1\{ga\})$	$(NP_2\{o\}\rightarrow\phi)$		*kōzi-suru*
3	$(NP_1\{ga\}\rightarrow\phi)$	$(NP_2\{o\}\rightarrow\{ga\})$	$(NP_3\{ni\})$	*tukeru/tuku*
4	$(NP_1\{ga\}\rightarrow\phi)$	$(NP_2\{o\}\rightarrow\{ga\})$	$(NP_3\{ni,e\})$	*syūka-suru*
5	$(NP_1\{ga\})$	$(NP_2\{o\}\rightarrow\phi)$	$(NP_3\{ni\}\rightarrow\{o\})$	*mukeru/muku*
6	$(NP_1\{ga\})$	$(NP_2\{o\})$	$(NP_3\{ni\}\rightarrow\phi)$	*sasou*
7	$(NP_1\{ga\})$	$(NP_2\{o\}\rightarrow\{kara,yori\})$		*dassyutu-suru*
8	$(NP_1\{ga\}\rightarrow\phi)$	$(NP_2\{o\}\rightarrow\{ga\})$	$(NP_3\{to,ni\})$	*setugō-suru*
9	$(NP_1\{ga\})$	$(NP_2\{ni\}\rightarrow\{o\})$		*hairyo-suru*
10	$(NP_1\{ga\}\rightarrow\phi)$	$(NP_2\{o\}\rightarrow\{ni\})$	$(NP_3\{de\}\rightarrow\{o\})$	*maku*

Table 1: The top-10 alternation types found in the valency dictionary

In Table 1, we present the top-10 scoring alternations extracted from the valency dictionary, which are then broken down into analytical, lexical and synthetic alternations in Table 2; for each of these, we present the score, number of alternation tokens and proportion of alternations for which some inconsistency (in selectional restrictions, lexical fillers or case marking) was detected.

Analytical alternations feature highly in the overall alternation token count, although lexical and synthetic alternations (particularly the passive and causative alternations) were found in reasonable numbers. The proportion of alternation tokens exhibiting some inconsistency is particularly high for lexical alternations, possibly due to the verb alternants being non-adjacent in the original dictionary, and hence more prone to error.

Next, we go on to evaluate the quality of the extracted alternations, based on a random sample of 260 alternation tokens from the 1,653 alternations actually incorporated into the dictionary. We classify each alternation token according to its type (analytical/lexical/synthetic), and then determine whether the candidate alternation is valid. In the case of a valid alternation, we distinguish between genuine alternations ("*OK*") and transfer dictionary quirks ("*Quirk*") such as adjunct optionality being described by way of two valency frames, with and without the adjunct. In the case of an error, on the other hand, we determine whether a genuine alternation does exist between the given case frames but was not identified correctly ("*X*"), or whether no plausible alternation exists ("*None*"). The statistical breakdown of these categories is presented in Table 3, with the raw count of each sub-category on the first row, and the

Index	***Total score***	***Analytical alts.***			***Lexical alts.***			***Synthetic alts.***		
		Score	***#***	***Inc't***	***Score***	***#***	***Inc't***	***Score***	***#***	***Inc't***
1	417.3	275.2	86	4.7%	118.9	34	22.1%	23.3	7	23.1%
2	127.8	118.6	54	6.0%	7.2	7	76.6%	2.0	2	13.3%
3	110.4	65.3	12	0.0%	26.2	11	54.1%	19.0	2	0.0%
4	97.7	45.6	12	10.9%	42.6	11	33.6%	9.6	1	0.0%
5	94.7	85.1	32	50.8%	9.6	2	100.0%	0.0	0	n/a
6	93.8	93.7	32	29.2%	0.1	2	100.0%	0.0	0	n/a
7	77.6	77.6	16	1.7%	0.0	0	n/a	0.0	0	n/a
8	74.4	74.4	8	4.9%	0.0	0	n/a	0.0	0	n/a
9	71.0	56.5	14	18.8%	14.5	3	16.6%	0.0	0	n/a
10	60.9	60.9	5	20.8%	0.0	0	n/a	0.0	0	n/a

Table 2: Statistical breakdown of the top-10 alternation types

Analytical alts.				*Lexical alts.*				*Synthetic alts.*			
OK	***Quirk***	**X**	***None***	***OK***	***Quirk***	**X**	***None***	***OK***	***Quirk***	**X**	***None***
90	19	13	21	30	0	4	26	17	1	1	39
.34	.07	.05	.08	.11	.00	.02	.10	.07	.00	.00	.15

Table 3: Evaluation of extracted candidate alternations

ratio of that count to the total number of candidate alternations in the second.

From Table 3, it is evident that genuine alternations and transfer dictionary quirks combined account for 60.9% of alternation candidates (meaning that roughly 1,007 of the 1,653 alternations in the final lexicon are valid). This constitutes an approximation of the accuracy of the extraction method. The method was remarkably successful in correctly identifying case slot correspondences in the case that some alternation did exist, erring in only 6.9% of cases overall. The remaining 33.0% of candidate alternations were postulated between case frames for which no correspondence exists. A significant number of these were found to be semantically related, but the nature of the semantic correspondence too vague to warrant an alternation analysis.

Only one instance of lexical filler variation was detected (where the one additional lexical filler was in fact a spelling variant of the two shared lexical fillers). Case marker variation was more common, and detected in 21 of the 261 manually-annotated candidate alternations. Out of these 21, 14 were contained in what were judged to be "correct" alternations, all of which were found to be instances of annotational inconsistency.

Selectional restrictions were preserved under alternation for 152 of the 261 annotated candidate alternations, of which 110 (72.4%) were found to be correct. Of the candidate alternations which did not preserve alternation, 49 were correct and 60 were incorrect (either the wrong mapping or a non-alternating valency frame pair). For most of the correct alternations where selectional restrictions were not preserved, the level of mismatch was slight (unsurprisingly given our scoring method). It remains to be determined whether these are annotational inconsistencies, transfer-specific variation or justified monolingual variation.

Note that there is a reasonable correlation between the score for a given candidate alternation and the quality of that alternation, such that increasing the alternation cutoff score from 0 could be used to bump up the overall accuracy, although at the expense of recall.

4 Representation in the lexicon

Having extracted the alternations, we need some way of describing them within the lexicon with minimal redundancy. We follow Baldwin et al. (1999) in employing a hierarchical structure comprising, in descending order, the `word` and `sense` levels.

The word level

At the topmost level, we describe features inherent to the lexeme in question, irrespective of its case frame realisation. These include its orthography, stem, reading and conjugational class. Importantly, we also describe any lexical and morphologically-unpredictable synthetic alternants of the lexeme here, classified according to alternation type. These take the form of links to other `word` entries. For *akeru*, for example, the inchoative alternant would be listed as a link to *aku*, whereas *kuru* "to come" would contain a listing for the morphologically idiosyncratic passive and synthetic causative forms of *koraeru* and *kosaseru*, respectively.

```
WORD:
    index = 958;
    orthography = ‘開ける’;  /* akeru */
    reading = ‘あける’; stem = ‘(開|あ) け’;
    part of speech = verb; conjugational class = v1;
    inchoative alternant = 959;

WORD:
    index = 959;
    orthography = ‘開く’;  /* aku */
    reading = ‘あく’; stem = ‘(開|あ)’;
    part of speech = verb; conjugational class = v5k;

SENSE:
    index = 958.3;
    VSA list = {(29/2),(16/2),(17/2)};
    alternation list = {causative-inchoative.lexical};

NP-ARG:
    index = 958.3.1;
    case-role = N1;   obligatory = False;
    selectional restrictions = {agent};
    lexical fillers = {};
    case marker list = {‘が’};  /* ga (nominative) */
    grammatical relation = subject; argument status = 3;

NP-ARG:
    index = 958.3.2;
    case-role = N2;   obligatory = False;
    selectional restrictions = {physical_object, facility};
    lexical fillers = {};
    case marker list = {‘を’};  /* o (accusative) */
    grammatical relation = object1; argument status = 3;
```

Figure 1: An example lexicon entry for *akeru/aku*

The sense level

Associated with each word is a set of senses. Each sense is defined to be a distinct alternation cluster, that is a set of case frames which can be derived directly (i.e. through a single alternation application) from a unique base case frame. The base case frame is described in the form of a list of arguments taken by it, and derived case frames are represented as a set of alternations which produce the desired case frame from the base case frame. The importance of our constraint on alternation direction becomes apparent at this point. As we only allow valency-decreasing and valency-maintaining alternations, the full argument set for a derived case frame is always going to be explicitly described for the base case frame, making it possible to transfer it directly across to the derived case frame.

Alternations are described by way of an alternation type (`causative-inchoative`, `unexpressed object`, `...`) and its associated alternation paradigm (`analytical`, `lexical` or `synthetic`). In the case of a lexical alternation, a list of links to lexical

alternants is provided at the `word` level, facilitating the retrieval of the correct verbal form. This can be seen in Figure 1 for the verb pair *akeru/aku*. Here, the case frame for the transitive *akeru* is the base form, which is associated with one alternation, the lexical causative-inchoative alternation. Therefore, the `word` entry for *akeru* provides a link to the (lexical) inchoative alternant *aku*, resulting in the derived unergative case frame being associated with *aku* as desired.

As described above, the extraction process has the potential to pick up on alternations where selectional restrictions and lexical fillers are not fully preserved. Eventually we hope to check all such anomalies and correct them where necessary. For now we are satisfied with a derived form of the lexicon which is informationally equivalent to the original valency dictionary. By default the selection restrictions and lexical fillers of the base case frame are copied across to the derived case frame. It is possible, however, to override such defaults by explicitly listing values for fields associated with the sense entry or case slots in the alternation description. This allows us to always recover the original verb entry.

Also represented at the sense level are verb semantic attributes (VSAs), a hierarchical description of the basic situation described by the verb (Nakaiwa et al. 1994). While any variation in VSAs under alternation is generally predictable, as with selectional restrictions and lexical fillers, it is possible to override the default value provided by the alternation mapping. This occurs within the alternation description.

Individual case slots contain an array of fields, including a case-role, grammatical relation, obligatoriness flag, selectional restrictions and lexical fillers.

5 Discussion

Here, we review research relevant to alternation extraction and application. Baldwin & Tanaka (2000) proposed a total of three alternation extraction procedures, using a full match or edge count-based similarity measure rather than entropy. Direct comparison of the different extraction methods is difficult, and Baldwin & Tanaka did not use the extracted alternations for any particular purpose. Schulte im Walde (2000) first derived subcategorisation frames with selectional restrictions, and then classified verbs into Levin (1993) classes according to the subcategorisation frames they occur with. Interestingly, she gained better results based on simple syntactic behaviour than when adding in selectional restrictions. McCarthy (2000) similarly derived subcategorisation frames with selectional restrictions from a corpus, using the minimum description length principle, and then classified verbs as participating in a select set of alternations according to similarity in selectional restrictions on corresponding case slots. One aspect of this research which sets it apart from that of both Schulte im Walde and McCarthy is that we compare selectional restrictions between corresponding case slots for arbitrary case frame pairings (for both the same and different verbs), rather than comparing corresponding case slots in equivalent frames across different verbs.

To summarise, this research is targeted at the derivation of a hierarchical, alternation-based lexicon from a flat-structured transfer dictionary. As part of this, we extracted alternations in an unsupervised manner, relying on the assumption that selectional restrictions are preserved under alternation. We proposed an entropy-based scoring method for evaluating both the degree of similarity and quality of match of a pair of selectional restrictions. This was used to score case frame mappings and analyse whether an alternation could be found between a given case frame pairing, a process which was found to be 60.9% accurate. The extracted alternation tokens were finally used in streamlining the lexicon by implicitly describing derived case frames as alternations over the base case frame.

Acknowledgements

This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank the three anonymous reviewers for their comments on the original version of this paper, and Sanae Fujita and Tomoko Kawaguchi for help with the alternation annotation.

References

Baldwin, Timothy, Francis Bond & Ben Hutchinson: 1999, 'A valency dictionary architecture for machine translation', in *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pp. 207–217.

Baldwin, Timothy & Hozumi Tanaka: 2000, 'Verb alternations and Japanese — how, what and where?' in *Proc. of the 14th Pacific Asia Conference on Language, Information and Computation (PACLIC 14)*, pp. 3–14.

Dorr, Bonnie J. & Mari Broman Olsen: 1996, 'Multingual generation: The role of telicity in lexical choice and syntactic realization', *Machine Translation*, **11**: 37–74.

EDR: 1995, *EDR Electronic Dictionary Technical Guide*, Japan Electronic Dictionary Research Institute, Ltd., (In Japanese).

Fujita, Sanae & Francis Bond: 2002, 'A method of adding new entries to a valency dictionary by exploiting existing lexical resources', in *Proc. of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2002)*, (this volume).

Fukui, Naoki, Shigeru Miyagawa & Carol 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.

Ikehara, Satoru, Masahiro Miyazaki, Akio Yokoo, Satoshi Shirai, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama & Yoshihiko Hayashi: 1997, *Nihongo Goi Taikei – A Japanese Lexicon*, Iwanami Shoten, 5 volumes. (In Japanese).

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, 'Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–', in *Proc. of the Third Machine Translation Summit (MT Summit III)*, Washington DC, pp. 101–106.

Jacobsen, Wesley M.: 1992, *The Transitive Structure of Events in Japanese*, Kurosio Publishers.

Levin, Beth: 1993, *English Verb Classes and Alterations*, University of Chicago Press.

McCarthy, Diana: 2000, 'Using semantic preferences to identify verbal participation in role switching alternations', in *Proc. of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2000)*, pp. 256–63.

Nakaiwa, Hiromi, Akio Yokoo & Satoru Ikehara: 1994, 'A system of verbal semantic attributes focused on the syntactic correspondence between Japanese and English', in *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, pp. 672–8.

Resnik, Philip: 1999, 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research (JAIR)*, **11**: 95–130.

Schulte im Walde, Sabine: 2000, 'Clustering verbs semantically according to their alternation behaviour', in *Proc. of the 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 747–53.