DALARNA
UNIVERSITY

Thesis

Bachelor's degree

The Past Tenses of Early Middle Japanese

Author: Arthur Hård
Supervisor: Mariya Aida Niendorf
Examiner: Herbert Jonsson
Subject/main field of study: Japanese
Course code: JP2011
Credits: 15 hp
Date of examination: January 18th, 2018

Abstract:

Early Middle Japanese is one of the oldest attested stages of Japanese. Its rich legacy consists of several literary works from the Heian era (7th to 11th centuries), some of which are still appreciated and widely read today. Despite a long tradition of research both within and outside Japan, quite a few details of the language remain incompletely understood. The present study addresses a long-standing question in the verbal domain of Early Middle Japanese, namely the semantics of the two so-called “past tenses” in *-ki* and *-ker-*. I tested the major hypotheses regarding their use by means of qualitative, corpus-based methods. Specifically, I trained a machine learning algorithm to predict which is likeliest of *-ki* and *-ker-* given a set of grammatical and semantic variables. Analysis of the results indicates that the suffixes likely embody a contrast between witnessed and non-witnessed past tense. It is also possible that mirativity—the grammaticalized expression of surprise at learning something unexpected—and aspect influence the choice of past tense suffix.

Keywords:
Early Middle Japanese, evidentiality, aspect, corpus linguistics

Contents

Conventions

Morpheme divisions and orthography are as in Tranter (2012) and may differ from other works on Early Middle Japanese. I use (Traditional) Hepburn for text names and common terms borrowed from Japanese.

Abbreviations of language and text names:

EMJ	Early Middle Japanese
GEN	Genjimonogatari
MS	Makura no Sōshi
OJ	Old Japanese
TM	Taketorimonogatari

Morpheme-by-morpheme glosses of Early Middle Japanese are in accordance with the Leipzig Glossing rules (available online at https://www.eva.mpg.de/lingua/resources/glossing-rules.php). Abbreviations:

0	semantically empty “linking” morpheme
ACC	Accusative
ADV	Adverbial
ALL	Allative
ATTR	Attributive
COM	Comitative
CONC	Concessive
DAT	Dative
DEB	Debitive
DS	Different Subject
EXCL	Exclamative
FIN	Final
FOC	Focus
GEN	Genitive
HUM	Humble
HON	Honorific
NEG	Negative
PERF	Perfect
PFV	Perfective
PREF	Prefix

POL	Politeness
SS	Same Subject
TENT	Tentative
TOP	Topic

I do not gloss *-ki* and *-ker-*.

When I cite examples from other authors, I sometimes modify their transcription scheme or interlinear in order to fit with the conventions of this thesis. All unattributed free translations into English are by me. Translations from other authors are taken verbatim from the original sources, with the exception of occasional clarifying material added by me (in brackets).

Chapter 1

Introduction

Early Middle Japanese (EMJ) is the common name for the literary Japanese of the Heian period (between the 7th and 11th centuries CE). Despite a long tradition of research both inside and outside Japan, many aspects of EMJ remain imperfectly understood. This study examines one particular blank spot in our knowledge, namely the semantics of the verb suffixes *-ki* and *-ker-*. Both are very common in Heian-era texts and they obviously performed an important function in the language. As such, examples are easy to find:

(1) *oni-no yau nar-u mono ide-k-i-te koros-am-u-to*
demon-GEN like be-ATTR thing exit-come-0-SS kill-TENT-CONC-COM
*s-i-**ki***
do-0-*ki*.FIN
"... demonic beings emerged and tried to kill us" (TM)

(2) *ima fa mukasi taketori-no okina-to if-u mono*
now TOP long.ago bamboo.cutter-GEN old.man-COM say-ATTR person
*ar-i-**ker**-i*
exist-0-*ker*-FIN
"A long time ago, there lived a person called Old Man Bamboo Cutter" (TM)

There is a relatively broad scholarly consensus that *-ki* and *-ker-* express past tense. Even those who consider EMJ to be a tenseless language (e.g. Vovin 2002 or Sandness 1991) concede that the suffixes often appear in passages that seem to refer to past time.

Since *-ki* and *-ker-* are mutually exclusive, it is highly likely that they stand in paradigmatic opposition to one another. Both appear in a wide range of lexical, pragmatic and grammatical contexts. In spite of ample data to work from, the ways in which the forms contrasted have proven elusive. Analyses in terms of evidentiality (e.g. Shinzato 1991) or aspect (e.g. Vovin 2002) are widespread and have some textual support, but the overall picture is one of disagreement among scholars.

The present work addresses *-ki* and *-ker-* from a novel perspective, informed by quantitative and corpus-based linguistics. I use machine learning to analyze the behaviors of the two suffixes across a set of EMJ texts, and forward an interpretation of their meanings that is in line with modern typological knowledge. The specifics of this procedure are explored in-depth in Section 3; it involves modeling the correlation between the occurrence of *-ki* and *-ker-* and other linguistic phenomena located at various levels of linguistic analysis (morphology, syntax, semantics and pragmatics). Although the purpose of this work is not methodological as such, I argue that quantitative and computational methods could be a powerful tool in future research on extinct languages such as EMJ.

The structure of this thesis is as follows: Chapter 2 provides a brief sketch of the EMJ verbal system before covering earlier research, 3 describes the methodology, while 4 contains an analysis of the results and further discussion and 5 finishes off the study with a summary.

Chapter 2

Background

2.1 The EMJ verb

EMJ is an agglutinative, suffixing language. Verbal morphology is rich but transparent, whereas nouns remain uninflected for the most part, case markers being clitics (Vovin 2002:47-48). Phonological adjustment across morpheme boundaries is rare, as are outright irregularities.

The structure of the EMJ verb is conveniently visualized according to a "slot-and-filler" model. The following schema is based on Vovin (ibid), with considerable modification:

Slot	**Class**
0	STEM
1	Honorification
2	Politeness
3	Negation
4	Aspect
5	Mood
6	*-ki*, *-ker-* and *-kem-*
7	Hearsay
8	Conjecture
9	Clause type

Table 2.1: EMJ verb structure

The stem in position 0 (or *P0*), which is sometimes internally complex, appears at the start of the word. Suffixes follow upon this, ordered according to the slots to which they belong; occasionally, the epenthetic segment *i* intervenes between a suffix and the site to which it attaches.

An example might be in place to demonstrate how polymorphemic verb forms can be broken down according to this model (the indices within parentheses point to slots; for the sake of clarity, epenthetic *i* has not been glossed):

(3) *obosimesi* *-tari* *-ker* *-unamer* *-i*
seem(P0) -CONT(P4) -*ker*(P6) -CONJ(P8) -FIN(P9)
"It seems like (Buddha) actually felt (pity for them)..." (*Uji Shūi Monogatari*, adapted from Tranter 2012:234)

Of course, a template is not sufficient to capture the full complexity of the EMJ verb, as certain combinations of grammatical categories surface as portmanteaus rather than the expected sequences of discrete morphemes. Furthermore, some slots are incompatible with each other. Such flaws notwithstanding, Table 2.1 is a useful reference.

The only slot that is obligatorily filled (i.e. always present) is P9, here entitled "Clause type." It houses a functionally diverse set of suffixes: modals (e.g. the Imperative in *-e* or *-yo*), nominalizers (e.g. the Attributive *-u* or *-uru*), negators (*-azi* "will not"), and various clause-linking devices such as *-edo* "but" and *-eba* "different subject converb."

The subjects of this paper, *-ki* and *-ker-*, are tentatively assigned to slot P6. *-ker-* is typically well-behaved and lines up with members of slots P7, P8 and P9 in typical agglutinative fashion. *-ki*, on the other hand, shows a very high degree of irregularity for EMJ, changing shape before or fusing with a subsequent suffix in P9. We can compare and contrast the behaviors of both forms before various P9 fillers:

Slot P9	*-ki*	*-ker-*
Final	-ki	-ker-i
Attributive	-si	-ker-u
Exclamative	-sika	-ker-e
DS Converb	-sika-ba	-ker-eba
Concessive	-sika-do	-ker-edo

Table 2.2: Allomorphs of *-ki* and *-ker-*

2.2 Previous research

EMJ has been the subject of centuries of research, both within and outside Japan. Consequently, a wealth of material exists on more or less every single facet of EMJ grammar, including monographs focusing solely on tense, aspect, and mood. Due to the difficulty of accessing publications written in Japanese, this section is limited to relatively recent publications in English.

The works discussed here do not necessarily cover the same time period as I do; in particular, some of the studies focus on more broadly on all forms of premodern Japanese (often under the label "Classical Japanese"), or take a diachronic or comparative approach and cite other stages of the language than EMJ. Nevertheless, all cover Heian-era Japanese to some extent, which motivates their inclusion in the present study. I also briefly address research *-ki* and *-ker-* in Old Japanese—which is temporally prior to EMJ and treated as a distinct linguistic entity in Tranter (2012)—in Section 2.2.5.

Authors also differ with regard to source text selection, which may well introduce bias in their conclusions. However, none of them provide a detailed breakdown of the evidence in favor of their hypotheses, which makes it difficult to say to which degree their analyses depend on any particular text. For this reason, I do not list the authors' sources unless it is highly relevant to do so.

Nor is extensive analysis of methodological issues in earlier accounts feasible. Many authors, including Vovin (2002), Sandness (1999) and Shinzato (1991), do not adhere to research procedures that are amenable to comparison. Takeuchi (1987) and Watanabe (2008) stand out in that they outline their theoretical priors and list the kinds of data they consider to be admissible evidence for or against an hypothesis. Since detailed examination of their methodologies would be beyond the scope of the present study, I will only bring it up if pertinent; the interested reader is advised to consult the source texts.

2.2.1 Takeuchi (1987)

Takeuchi (1987) argues that the core distinction encoded by *-ki* and *-ker-* is one of *evidentiality*, a grammatical category whose function is to express the spealer's source of information (Aikhenvald 2003:1).

According to Takeuchi's analysis, the primary function of *-ki* is to situate an event in the past, and further to imply that "the events in question have been observed directly by the speaker" or that "he or she believes or pretends to believe" that they actually happened (ibid, 21-22). *-ker-* is used whenever the speaker did not witness the event but knows of it through hearsay or inference (ibid, 22). These conclusions stand on distributional evidence. In particular, *-ki* preferentially occurs in first-person narration, such as

(4) *tosigoro* *wosana-ku* *faber-i-si-yori*
age immature-ADV be.POL-0-*ki*.ATTR-from

"(I have never left her) all these years ever since I was a child ..." (GEN, adapted from Takeuchi 1987:23)

(5) *sakizaki-mo* *maus-am-u-to* *omof-i-sika-domo*
before-FOC tell.HUM-TENT-FIN-COM tell-0-*ki*-CONC

"I intended to tell you before, but ..." (TM, adapted from Takeuchi 1987:24)

where a character (different in each case) is recounting their own experiences.

-ker-, in turn, often appears in descriptions of situations happening outside the narrator's consciousness, e.g.

(6) *kotosi-yori-fa* *futagar-i-ker-u* *kata-ni*
this.year-from-TOP be.tabooed-0-*ker*-ATTR direction-DAT
faber-i-ker-eba
be.POL-0-*ker*-DS

"... we learnt that it is in a direction which is tabooed starting this year" (GEN, adapted from Takeuchi 1987:23)

Presumably, the location of the place, and the fact that the direction was tabooed, are known to the speaker only through secondhand means ("we learnt").

Two-term evidentiality systems of the proposed kind are very common throughout Eurasia (Aikhenvald 2004:15). Typologically, it would not be unusual to encounter such a contrast in EMJ.

2.2.2 Shinzato (1991)

Shinzato's (1991) account is largely similar to Takeuchi (1987), although she goes into more detail and ascribes a number of aspectual and non-evidential "epistemic" senses to both forms.

Pointing to the distributional facts, Shinzato claims that *-ki* "conveys information which was acquired through the speaker's direct experience" and that *-ker-* "entails information which the speaker infers ... or hears second-hand" (ibid 35-36). In addition to these uses, both forms have a number of senses which, at face value, seem to contradict the evidentiality theory. Specifically, there is a marked preference for *-ki* in historical accounts, which normally do not refer to events which the narrator perceived, as well as in Japanese translations of Chinese documents regardless of their contents (ibid, 38-39). Note that Shinzato's examples are from the *Kojiki*, which is an Old Japanese text; I nevertheless reproduce one of them to illustrate the point.

(7) *kare sono usagi yasokami-no osife-ni sitagaf-i-te*
so that rabbit Yasokami-GEN advice-DAT follow-0-SS
fus-i-ki
lie.down-0-*ki*.FIN

"So that rabbit, following Yasokami's advice, lay down" (*Kojiki*, adapted from Shinzato 1991:38)

Another deviation is *-ker-* as a marker of surprise (or "mirativity," cf. DeLancey 1997):

(8) *syari-wo todome-tamaf-u koto-fa syuzyaunomi-wo*
bones-ACC keep-HON-ATTR thing-TOP living.things-ACC
eki-s-i-tamaf-am-u-to nar-i-ker-i
benefit-do-HON-TENT-FIN-COM be-0-*ker*-FIN

"... keeping (Nyoraitaishi's) bones would widely benefit living things" (*Konkōmyōsaishō-ookyō*, adapted from Shinzato 1991:39)

The sentence is framed by a longer statement to the effect "I just realized that..."; the final predicate *nar-i-ker-i* could perhaps be translated as "it turns out that." The use of the past tense in statements of surprise is common in contemporary Japanese, but it is telling that whereas such usages occur for EMJ *-ker-*, they are seemingly absent for *-ki*.

Shinzato (1991:42-32) attempts to provide a unifying explanation for the various functions of *-ki* and *-ker-*, proposing that the contrast between them has to do with "epistemicity," that is, with "conveying integrated vs. non-integrated information." The argument rests on a typological comparison with modern Turkish, which likewise has two past tense forms that distinguish between directly and indirectly experienced events (Aksu-Koç and Slobin 1986:159-167). Due to universal trends of grammaticalization, it is not altogether implausible for two otherwise unconnected languages to exhibit a significant degree of similarity in certain grammatical domains (Andrason 2016).

2.2.3 Sandness (1999)

Unlike the other authors cited in this section, Sandness (1999) does not believe that *-ki* and *-ker-* are a pair. Instead, they are historically unrelated suffixes that perform widely divergent functions.

Sandness maintains that *-ki* is comparable to the imperfect tense of Romance languages, "[t]hat is, it refers to habitual past action, action taking place over a long period of time, or past action serving as a background for the main action of the narrative" (ibid, 41). She presents many examples suggestive of such an interpretation, like:

(9) *fanazakura* *sak-u-to* *mi-si* *ma-ni* *katu*
cherry.blossoms bloom-FIN-COM watch-*ki*.ATTR period-DAT yet
tir-i-n-i-ker-i
scatter-0-PFV-0-*ker*-FIN

"As *I was watching* the cherry blossoms, at the same time, they seemed to start scattering" (*Kokinshū*, adapted from Sandness 1999:40, emphasis original)

(10) *inisife-ni* *ar-i-ki* *ar-az-u-fa* *sir-an-edomo*
past-DAT exist-0-*ki*.FIN exist-NEG-FIN-TOP know-NEG-CONC

"I don't know whether there *used to be* such things in the past or not" (*Kokinshū*, adapted from Sandness 1999:40, emphasis original)

-ker- is a "subjectivizer," a kind of modal that is not specifically connected to past time (ibid, 54). It signals that the speaker is present his or her own perception of an event, which explains the choice of *-ker-* rather than *-ki* in similes and other types of poetic language. Thus, in

(11) *nami-no* *ut-u* *se* *mi-reba* *tama-zo* *midare-ker-u*
wave-GEN beat-ATTR rapid see-DS jewel-FOC disordered-*ker*-ATTR

"When I see the rapids where the waves beat, jewels *are scattered* on them" (*Kokinshū*, adapted from Sandness 1999:50, emphasis original)

Sandness (ibid, 51) explains the choice of *-ker-* as a strategy to cast the statement as a metaphor, noting that "[i]n real life ... jewels do not float on rapids." Surprise and hearsay tokens of *-ker-* should be taken as special cases of this core meaning.

2.2.4 Vovin (2002)

Vovin (2002) considers EMJ to be a tenseless language, i.e. to lack tense as a grammatical icon. Rather, he groups *-ki* and *-ker-* as "retrospectives," which denote the speaker's recollection of facts without any reference to time. Practically, the vast majority of tokens of the suffixes do refer to the past, but that is incidental and not a part of their core semantics.

Echoing Takeuchi (1987) and Shinzato (1991), Vovin defines *-ki* as an "subjective retrospective," which "expresses recollection of an event directly experienced by the speaker" (ibid, 225). The evidence for this is distributional: in fiction, *-ki* is very common in the dialogue of characters who are recounting their own experiences. However, usages that deviate from this pattern are not uncommon, e.g.

(12) *konron yama-ni ir-i-tamaf-i-n-i-ki*
Kunlun mountain-DAT enter-0-HON-0-PFV-0-*ki*.FIN

"(the Emperor) entered the Kunlun mountains" (*Hamamatsu chūnagon monogatari*, from adapted from Vovin 2002:288)

(13) *tukafas-i-si fito-fa yoru firu mat-i-tamaf-u-ni*
send-0-*ki*.ATTR people-TOP night day wait-0-HON-ATTR-DAT

"(the Dainagon) waited day and night for the people whom (he) had sent, but ..." (TM, adapted from Vovin 2002:227)

The second example is from *Taketorimonogatari* and is a part of the main narrative, in which *-ker-* is otherwise dominant. According to Vovin, the context of the first sentence also makes it clear that the narrator did not see the Emperor entering the mountains.

-ker- is an "objective" retrospective, "a marker that expresses recollection of an event about which he has received information from someone else" (ibid, 304). It also appears in statements of surprise. This is in line with the evidentiality/mirativity hypothesis investigated in earlier section. As an alternative hypothesis, Vovin observes that *-ker-* could plausibly indicate imperfective or progressive aspect (ibid, 304-305), in addition to or instead of any evidentiality-like functions.

2.2.5 Evidence from Old Japanese

-ki and *-ker-* existed already at the earliest recorded stage of Japanese, which is conventionally termed Old Japanese (or OJ). Diachronically, OJ directly precedes and develops into EMJ, so it might be worthwhile to spend a few words on its counterparts to *-ki* and *-ker-* (more properly *-kyer-*, cf. Bentley 2012:204).

Reminiscent of Vovin (cf. 2.2.4), Bentley (2001:156-158) calls *-ki* a "perfective retrospective." Confusingly, he does not explain exactly what this label means, but as he adds the tag "[**we know**]" to the end of some of his translations of sentences containing *-ki*, it should probably be interpreted as a kind of direct evidential. Nor does he discuss *-kyer-*; however, in Bentley (2012:191-211), he refers to it as a second retrospective, without going into any detail.

Frellesvig (2010:72-76) treats *-ki* as a simple past, used for historical as well as directly experienced events. The "Modal past" *-kyer-* is more elusive, and seems to have aspectual, modal and mirative functions. Eventually, he settles on a characterization similar to that of Sandness (cf 2.2.3) and says that it "expresses addresser involvement and subjectivity."

Watanabe (2008:91-127) interprets *-ki* and *-kyer-* as analogous to the modern French *passée simple* and *imparfait*. The contrast would thereby be one of perfective versus imperfective aspect. The evidence in favor of this are partly diachronic, as Watanabe claims that *-kyer-* originates from a longer construction involving the suffix *-yer-* which has imperfective semantics. She rejects evidentiality and mirativity, claiming that such readings arise due to a confusion of the form and "the pragmatic context where it is used" (ibid, 125).

2.2.6 Summary

Table 2.3 presents a summary of the views covered in this section.

Author	**Stage**	**Summary**
Takeuchi 1987	EMJ	Direct versus indirect evidentiality, aspectually neutral
Shinzato 1991	EMJ	Evidentiality, mirativity and aspect
Sandness 1999	EMJ	*-ki* is a past with imperfective semantics, *-ker-* is a subjectivizer
Vovin 2002	EMJ	Direct versus indirect recollection of an event, *-ker-* might be imperfective, progressive, or resultative
Bentley 2001, 2012	OJ	Presumably direct versus indirect recollection
Frellesvig 2010	OJ	*-ki* is a past tense, while *-kyer-* denotes speaker involvement and subjectivity
Watanabe 2008	OJ	Perfective/punctual past versus imperfective past

Table 2.3: Summary of previous research

The majority of researchers invoke the notions of evidentiality, aspect and/or mirativity to explain the apparent contrast between *-ki* and *-ker-*, with only Sandness (1999) disagreeing substantially.

Chapter 3

Methodology

3.1 Introduction

To reiterate, the objectives of this study are

1. to test how well previous accounts of *-ki* and *-ker-* line up with distributional evidence (which I do using a machine learning model), and
2. to propose novel interpretations if data support for such turns out to exist.

The general approach I adopt to accomplish these goals is similar to the corpus-linguistic analysis method *behavioral profiling*. The central insight is that a correlation exists between the distribution of a linguistic entity and its semantics (Divjak and Gries 2006). Much of the research utilizing this principle focuses on near-synonymy, with the aim to find out the subtle ways in which seemingly identical words or constructions differ (e.g. Arppe 2008). There are fewer studies which tackle questions of grammatical meaning with it, as this one does, but no there is no reason for why it would not be suitable.

Very briefly explained, behavioral profiling and similar procedures consist of three stages:

1. creation of a set of samples (= "data set") of the linguistic phenomenon under consideration
2. the annotation of each sample according to a range of linguistic properties that are deemed relevant to the study (= "variables" or "dimensions"), and
3. application of statistical modeling to assess the degree of correlation between the variables and the occurrence of the phenomenon.

See Gries (2010) for a more detailed exposition.

I will address the execution of these steps out of order, beginning with the modeling (3) in Section 3.2. Unlike the majority of corpus studies, I use machine learning rather than statistics; additionally, the scope of my work is slightly different from the mainly lexicographic research that inspired it (e.g. Divjak and Gries 2006, Arppe 2008, Jansegers and Gries 2017 and others). With this out of the way, I turn to (1) data collection 3.3 and annotation 3.4, which are perhaps the most vital parts of the project.

The annotated source data is available online at https://github.com/aehard/jpdata (Hard 2019).

3.2 Modeling with machine learning

There is a growing recognition, thanks in part to recent advances in computational linguistics, that linguistic choices often have a plurality of causes (Arppe 2008:7-11). Grammatical, contextual, sociological, pragmatic and neurological factors all conspire to determine the outcome of decision processes, and it can sometimes be hard or impossible to pinpoint a single motivating factor. Accurate modeling of linguistic alternatives requires taking a whole range of factors into consideration, and determining the degree of influence each yields over the outcome. One way to do this is statistics (e.g. in Poplack 1992, Arppe 2008 and others), but in this study I rely on the related field of machine learning.

Broadly speaking, machine learning is a discipline concerned with the creation of computer programs that are able to learn how successfully execute some task (Samuel 1959). Many prototypical machine learning problems involve prediction, inference and classification: identifying humans by their facial features, predicting market directions from the fluctuations of economic indicators, or determining attitudes towards a company from online product reviews are only some examples. In most cases, success is a function of data quantity and quality rather than the particular learning algorithm (Redman 2016).

I treat the problem of *-ki* and *-ker-* as a classification task. The challenge is to decide which form is more appropriate in a given context, as defined by various contextual variables. The classifier learns to do this by assigning "weights" to the variables, emphasizing those that are highly reliable cues for suffix choice while de-emphasizing those that are uncorrelated with it. Baayen et al (2013) demonstrates that variable weighting by means of a machine learning algorithm is potentially equivalent to logistic regression for determining the contribution of each factor in linguistic modeling.

Following the recommendation in Baayen et al (ibid), I train a random forest algorithm (Ho 1995, Breiman 2001) on EMJ data. The label "random forest" accurately describes what they are: ensembles of decision trees which pool their results together, returning the majority result as the output (such as the class to which a data point is assigned). Figure 3.1 shows a hypothetical individual tree.

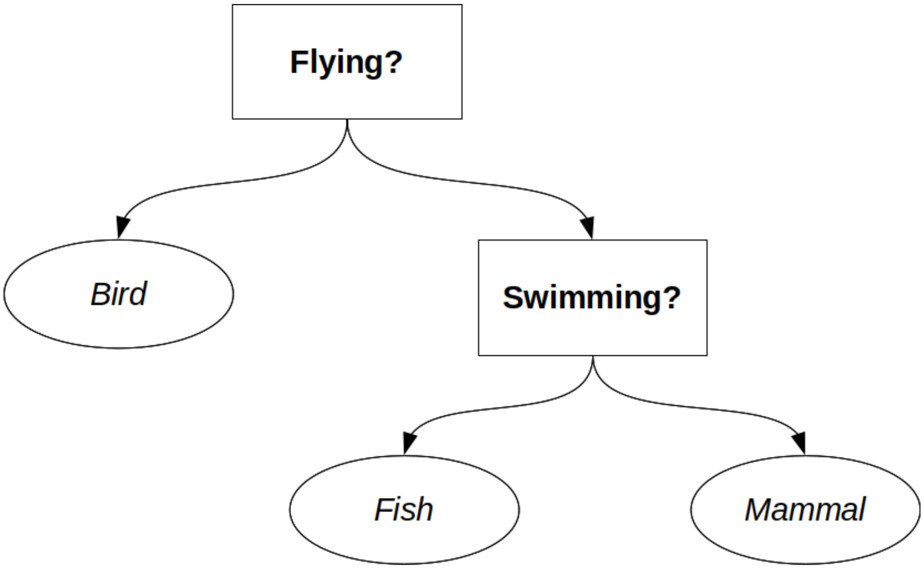

Figure 3.1: Sample tree in a hypothetical random forest. Leaf nodes are outputs; left path from an internal node represents “yes” to the question, and right path “no.”

A typical random forest consists of a multitude of such trees, in which the bifurcations correspond to a randomly chosen subsets of model features. Figure 3.1 forks at FLYING? and SWIMMING?, but other trees in the same hypothetical forest may use other variables (e.g. CRAWLING?, JUMPING? and so on) at its branching sites. Training involves tuning each tree to minimize erroneous outputs.

Random forests are powerful and enjoy considerable popularity in analytics and Big Data circles (Denil et al 2014). Consequently, there are many good, freely available resources for implementing them. Another benefit more tangibly related to the present study is that it is relatively simple to calculate variable weights. The particular library I utilize, the randomForest library in R (Liaw and Wiener 2002), measures this in terms of the *mean decrease in Gini-gain* incurred by randomly “scrambling” the variable values. As an example, assume that a model contains a factor VEHICLE TYPE with two levels of depth, “car” and “boat,” and that the trained random forest learns to associate the former with one price class and the latter with another. It would in that case be possible to calculate the degree of contribution of VEHICLE TYPE as the overall loss of performance that results from permuting its values across the forest, i.e. changing “car” into “boat” and vice versa.

Importance scores enable rank-ordering of the variables by their impact on the classification task. Those with a high rank (1 being the highest) have bigger effect on the outcome; conversely, low rank implies that the outcome is largely independent of that variable. Ranking thereby serves as a viable alternative to statistical significance testing, such as calculating p-values.

3.3 Data acquisition

The data comes from three texts, namely the entirety of *Taketorimonogatari*, parts of *Makura no Sōshi* and some pargraphs of *Genjimonogatari*. The basic information and number of data points extracted from each is given in Table 3.1.

Text	**Original author**	**Edition**	*-ki*	*-ker-*
Taketorimonogatari	unknown	Takeda (2001)	33	74
Makura no Sōshi	Sei Shōnagon	Sakaguchi (2001)	41	24
Genjimonogatari	Murasaki Shikibu	(online version)[1]	9	2

Table 3.1: Source text editions; the columns *-ki* and *-ker-* hold counts of those suffixes

Practical concerns guided text selection: availability of modern editions, and availability of translations into contemporary Japanese. The doubtful sourcing of the *Genjimonogatari* edition prompted me to largely exclude it, and only a few data points from it made their way into the final study. Time concerns prevented me from localizing a better source.

Gries (2010) recommends drawing a random set of tokens of the item under investigation, presumably to minimize sampling bias. Unfortunately, Japanese tends to drop established information from discourse, so that it can be very hard to correctly interpret a sentence without knowing its larger context. As this would create problems for annotation, I have instead opted for simply reading through the texts, collecting any instances of *-ki* and *-ker-* encountered along the way.

3.4 Annotation

I briefly touch upon the importance of data advantage for machine learning in Section 3.2. Having a large amount of good data is ideal, but in practice developers often opt for either quality or quantity. the present study relies on a limited set of high-quality samples for practical reasons, as the data acquisition procedure is extremely time-consuming.

Quality is a vast topic within machine learning. The algorithms can sometimes discover useful features from raw data; nevertheless, it is often advisable to "skew" the learning process by forcing the data into a representation that emphasizes relevant features, or deemphasizes irrelevant ones. Case in point: many statistical document classifiers reduce documents to unordered lists of content words ("bags of words") by removing all function words (Manning and Schütze

[1] URL: http://jti.lib.virginia.edu/japanese/genji/index.html

1999:235). Such “bags” are less information-rich in comparison to the original texts, but facilitate automatic classification as the information that does remain is more useful and more accessible.

For the present study, quality means zooming in on the features that are likely to be relevant for the choice between *-ki* and *-ker-*. For instance, it is probably safe to reject “number of nouns in the sentence” as a possible contributing factor out of hand. On the other hand, the EMJ literature gives a fair number of suggestions, mostly revolving around the grammatical, lexical or pragmatic context, that are almost certainly relevant.

It would be desirable to use as many as possible of these potential explanatory factors in the model (cf. Arppe 2008:28-32). Unfortunately, large variable spaces may inadvertently lead to overfitting, which is what happens when a machine learning model becomes “too good” on a single data set while failing to generalize. Furthermore, some of the proposals found in earlier accounts are not amenable to operationalization. For instance, Sandness’s (1991) claim that *-ker-* is frequent in metaphors is very hard to test, as there are no obvious formal criteria for judging whether a statement is metaphorical or not. To avoid introducing an undue amount of subjectivity, most of the variables I include in the model have strictly formal definitions.

The final variable set has nine members, many of which can take multiple values, as tabulated in Table 3.2.

Variable name	**Values**
Co-occurring Suffix	none, *-tar-/-er-*, *-t-/-n-*, prospective construction
First Person Subject	true, false
Personal Experience	true, false
Syntactic Context	final, kakarimusubi 1, kakarimusubi 2, converb in *-eba*, converb in *-edo*, subject relativization, non-subject relativization, empty noun relativization, juncture with =*ni*, juncture with =*wo*, other juncture, nominalization
Embedding	none, perception embedding, cognitive embedding, other
Lexical Verb	(lexical verb)
Lexical Type	adjective, intransitive verb, transitive verb
Bounded	true, false
Present Reference	true, false

Table 3.2: Variables and their levels of depth

I will in turn cover each variable’s definition and the rationale behind its inclusion.

3.4.1 Co-occurring Suffix

co-occurring suffix codes the presence or absence of certain suffixes on the verb, all of which deal with the semantic category of *aspect*.

Level	**Assignment criteria**
none	default
perfect	*-tar-* or *-er-* in P4
perfective	*-t-* or *-n-* in P4
prospective	verb in the frame verb-*am-u-to s-*

Table 3.3: Levels of depth for Co-occurring Suffix

The suffixes *-tar-* and *-er-* are functional equivalents and express "continuation of a state resulting from an earlier action" (Tranter 2012:232), whence my grouping of them under the label *perfect*. I pair up *-t-* and *-n-* as for similar reasons, as many researchers suspect that they are lexically conditioned allomorphs (ibid, 232). Finally, EMJ has a fairly common construction, verb-*am-u-to s-*, which is comparable in meaning to contemporary Japanese verb-*yoo to suru* ("about to do"). Although not a suffix, I nevertheless code it as a value of co-occurring suffix.

-ki and *-ker-* often combine with aspect-marking suffixes in EMJ texts, which might indicate that they are underspecified with regard to aspect, as Takeuchi (1987) argues, or that the language simply allows combinations of forms with different aspectual values. In either case, grammatical aspect is clearly related to the central problem of characterizing the semantics of *-ki* and *-ker-*, and co-occurring suffix is an easy way to operationalize this dimension.

3.4.2 First Person Subject

first person subject receives the value *true* when the speaker is the grammatical subject of the verb, and *false* otherwise. Some doubt exists as to the morphosyntactic type of OJ (see e.g. Vovin 1997), but it seems reasonable to call EMJ a nominative-accusative language. In other words, it is possible to define the subject as the nominative-marked argument of a verb. However, EMJ only has explicit marking of the nominative case in certain clause types (Akiba 1978:112-121), which makes it hard to find a good formal definition.

To get around this problem, I rely on a number of heuristic rules for identifying the subject of a predicate:

1. the constituent marked by the Nominative case "particles" *-no* and *-ga*

2. the most agent-like or experiencer-like argument of a transitive verb

3. the sole argument of any intransitive verb, including passivized verbs

4. the first argument in copula constructions of the type *X Y nar-i* "X is Y"
5. the first argument in so-called "double subject" constructions

FIRST PERSON SUBJECT touches on evidentiality. Interaction between grammatical person and evidentiality is fairly widespread; see Curnow (2001) for a review.

3.4.3 Personal Experience

PERSONAL EXPERIENCE is likewise a binary variable. It counts as *true* when the speaker or narrator directly perceived the event, whether visually or through some other sense. Hearsay, conjecture, counterfactual statements, and "narrative" events in tales like *Taketorimonogatari* or *Genjimonogatari* receive the value *false*. In ambiguous cases, I err on the side of caution and assign *false*.

3.4.4 Syntactic Context

With its 14 levels, SYNTACTIC CONTEXT is the most complicated variable in the model. It codes the overall grammatical context of the central verb of the data point.

EMJ is strictly verb-final. Broadly speaking, EMJ clauses fall into three categories ("types") according to their functional and formal properties.

1. sentence-final clauses
2. adnominal clauses, which are always in the Attributive and form relative clauses
3. cosubordinated clauses, which are coordinated with or subordinated to the subsequent clause

With this in mind, we can state the definitional criteria for the 14 values of SYNTACTIC CONTEXT:

Value Assignment criteria	
Final	Clause is sentence-final and verb ends in Final form
Kakarimusubi 1	Clause is sentence-final, verb ends in Attributive form, and one of the particles *zo, namu, woba* occur earlier in the sentence
Kakarimusubi 2	Clause is sentence-final, verb ends in Exclamative form, and the particle *koso* occurs earlier
Subject Relative	Clause is adnominal and modifies its subject
Other Relative	Clause is adnominal and modifies an argument that is not its subject
-*eba* Converb	Clause is cosubordinated and verb ends in -*eba*
-*edo* Converb	Clause is cosubordinated and verb ends in -*edo*
Clausal Juncture	Clause is cosubordinated, verb ends in Attributive form, and is followed by a case marker
Nominalization	Clause is in Attributive form, optionally followed by a semantically empty noun like *koto* "thing," and functions as a noun in its own right

Table 3.4: Overview of SYNTACTIC CONTEXT

For in-depth coverage of these various constructions, the reader is advised to consult Akiba (1978) for EMJ syntax in general, Ohori (1992:128-181) for cosubordination ("clausal juncture"), and Ohori (ibid, 181-210) for relative clause formation. Here, I will only briefly describe some of the unfamiliar terms.

Most sentence-final clauses in EMJ end in a verb inflected for the Final form. Occasionally, the last verb in a sentence bears the Attributive or Exclamative instead, which signifies a kind of focusing cleft construction known as *kakarimusubi* in the Japanese grammatical tradition. My translation of the example below is admittedly somewhat artificial but captures roughly the pragmatic effect kakarimusubi seems to have:

(14) *ima akikaze fuk-am-u wori-zo ko-m-u-to s-uru*
now autumn.wind blow-TENT-ATTR time-FOC come-TENT-COM

do-ATTR

"It is when the autumn winds blow that I will return" (MS)

The various "cosubordinate" constructions are similar to English subordinators and conjunctions (e.g. "and," "but," "if" and so on), e.g.

(15)	*yo-u*	*kakus-i-tar-i-to*	*omof-i-si-wo*
	well-ADV	hide-0-PERF-FIN-COM	think-0-*si*.ATTR-ACC

"... although I thought I hid it well" (MS)

Here, the Attributive plus the case marker *-wo* Accusative correspond roughly to English "although."

Marking of grammatical categories such as evidentiality, aspect and mood often exhibits constraints and preferences with regard to syntactic environment (Aikhenvald 2004:241-270). This makes SYNTACTIC CONTEXT a potentially useful clue for the semantics of *-ki* and *-ker-*.

3.4.5 Embedding Context

The variable EMBEDDING CONTEXT codes the semantic type of the controlling verb whenever the central verb of the data point appears as part of a complement clause. Specifically,

Value	**Criteria**
none	verb is not complementized
perception	controlling verb denotes sensory perception: vision, hearing, taste, tactile sensations and so on
cognition	controlling verb denotes thinking, emotion, reminiscence or similar mental process
other	any other kind of controlling verb

Table 3.5: Overview of EMBEDDING CONTEXT

The commonest strategy for creating complement clauses in EMJ uses the case marker or "particle" *-to* (Akiba 1978:89-92). Other strategies also occur, such as nominalization by means of the Attributive form, or relative clauses headed by the semantically empty noun *koto* "thing." The details of formulation are irrelevant for this variable, which only references the higher verb.

3.4.6 Present Reference

The binary variable PRESENT REFERENCE codes whether the data point has connotations of realization. Formally, when a speaker refers to a fact as if he or she had not known it before, this variable becomes *true*.

3.4.7 Bounded

The variable BOUNDED is binary. *true* implies that the event encapsulated by the data point has a beginning and an end. The typical cases of a bounded action is physical activity that is limited in time and space, e.g.

(16) *mina nanifa-made miokuri s-i-ker-u*
all Naniwa-TERM farewell do-0-*ker*-ATTR
"... all (who worked for him) saw him off in Naniwa" (TM)

The farewell-taking is a single occasion with clear boundaries. Compare this with

(17) *koko-fe-to simo omof-azar-i-ker-u fito-mo*
here-ALL-COM even think-NEG-0-*ker*-ATTR people-FOC
... "even people who did not think (they would arrive) here" (MS)

which describes an event ("think") that is not localized or bounded in any obvious way. The fact that the verb *omof-* is negated means that it has neither a beginning nor an end, and furthermore the situation depicted in the sentence involves different people experiencing the same confusion, presumably over an undefined period of time.

Repetition of events with boundaries also assign the value *false* to BOUNDED, such as

(18) *noyama-ni mazir-i-te take-wo tor-i-tutu yorodu-no*
mountains-DAT enter-0-SS bamboo-ACC take-0-WHILE myriad-GEN
koto-ni tukaf-i-ker-i
thing-DAT use-0-*ker*-FIN
"He (= Old Man Bamboo Cutter) went into the mountains, retrieved bamboo and used it for a myriad of things" (TM)

which I assume denotes a habitual activity as it is the profession of Old Man Bamboo Cutter.

The temporal structure of events, often called *aktionsart*, is closely related to aspect as a grammatical category. There are many ways to classify verbs according to their aktionsart, although the most common is perhaps that of Vendler (1957), which divides actions into *accomplishments*, *achievements*, *activities* and *states* according to duration and presence or absence of a natural endpoint. Grammatical marking of tense and aspect interact with such semantic features in various ways; a good example is how the *-te iru* construction in contemporary Japanese takes on a resultative reading with *accomplishments* and *achievements*, and a progressive reading with *activities* (Tranter and Kizu 2012:291).

The Vendlerian scheme is not ideal for the purposes of the present study, despite its widespread use. The problem is that deciding the Vendlerian class of a particular verb hinges on being able to subject that verb to a variety of co-occurrence "tests." Doing this for EMJ would be a reseach project in its own right; instead, I have opted to rely on the much simpler scheme outlined above.

Lexical Verb ID and Lexical Verb Type

The variable LEXICAL VERB ID holds the verb to which *-ki* or *-ker-* attaches, sans any inflections. Compound or serial verbs count as single lexical items (cf. Lanz 2009).

LEXICAL VERB TYPE codes whether the host verb is transitive, intransitive or an adjective verbalized by the suffix *-kar-*. A fourth possible value is *copula*, assigned whenever the host verb is *nar-* "be" or some variant thereof.

Chapter 4

Results

4.1 Analysis

4.1.1 Variable importance

I used the randomForest library (Liaw and Wiener 2002) for the statistics and machine learning software R (R Core Team 2013) to train a random forest with these specifications (shown in Table 4.1).

Parameter	**Value**
Tree count	100
Seed	123
Training data	90
Test data	92

Table 4.1: Random forest parameters

Note that randomForest does not allow variables with more than 57 levels of depth, forcing me to discard LEXICAL VERB ID. Additional testing with the random forest implementation in the common lisp machine learning library clml, which does tolerate it, did not yield substantially different results.

As the confusion matrix in Table 4.2 shows, performance was fairly good:

	-ki	***-ker-***	Error percentage
-ki	25	12	32.4%
-ker-	9	49	15.5%

Table 4.2: Confusion matrix. Top row is actual class, left column predicted class

The total error rate (correct classifications divided by total classifications) is 22.11%, meaning that the model correctly guesses the appropriate past tense marker nearly four times out of five. While there is certainly room for improvement, these results indicate that the model is on the right track.

To understand the inner workings of the model, it is necessary to estimate the importance of each individual variable. The normal way to do this with a random forest is to arbitrarily change the values of the variables and measure the performance decrease this incurs (Louppe et al 2013). Note that these numbers reflect *average* contribution. A variable with a low score, such as EMBEDDING, may be predictive under specific conditions, but nevertheless be of little utility overall.

Table 4.3 and Figure 4.1 give the importance scores for each of the variables.

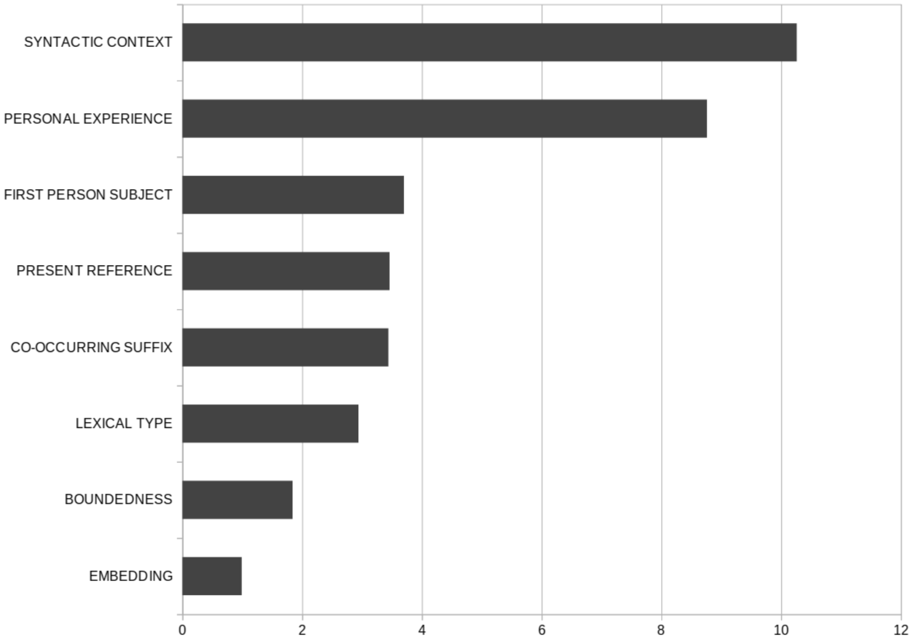

Figure 4.1: Ranked variable importance by mean decrease in Gini-gain

Variable	**Importance**
Syntactic Context	10.26
Personal Experience	8.76
First Person Subject	3.7
Present reference	3.46
Co-occurring Suffix	3.44
Lexical Type	2.94
Boundedness	1.84
Embedding	0.99

Table 4.3: Ranked variable importance estimated by mean decrease in Gini-gains

The second-best predictor is personal experience, which seems to confirm the evidentiality hypothesis. On the other hand, the aspect-related variables co-occurring suffix and boundedness fare rather poorly, perhaps implying that aspect has little effect on the choice between *-ki* and *-ker-*.

4.1.2 Visualization

The variable importance scores are useful mainly as indicators of the impact of individual factors on past tense choice. Since linguistic phenomena often have multiple causes (Arppe 2008), it would be interesting to also look at non-linear interaction effects among the variables. Unfortunately, the ensemble nature of random forests makes it hard to do so. No one tree represents the decision-making process embodied by the entire forest; selecting an arbitrary tree to analyze will not yield any understanding.

As a workaround, I trained a single decision tree (with the R Core package) on the same data as the random forest, under the assumption it would arrive at a comparable level of performance while being easier to plot. It is depicted in Figure 4.2. The number of leaf nodes was set to 5, as performance gains beyond this point were minimal and probably resulted in overfitting (cf. Quinlan 1986).

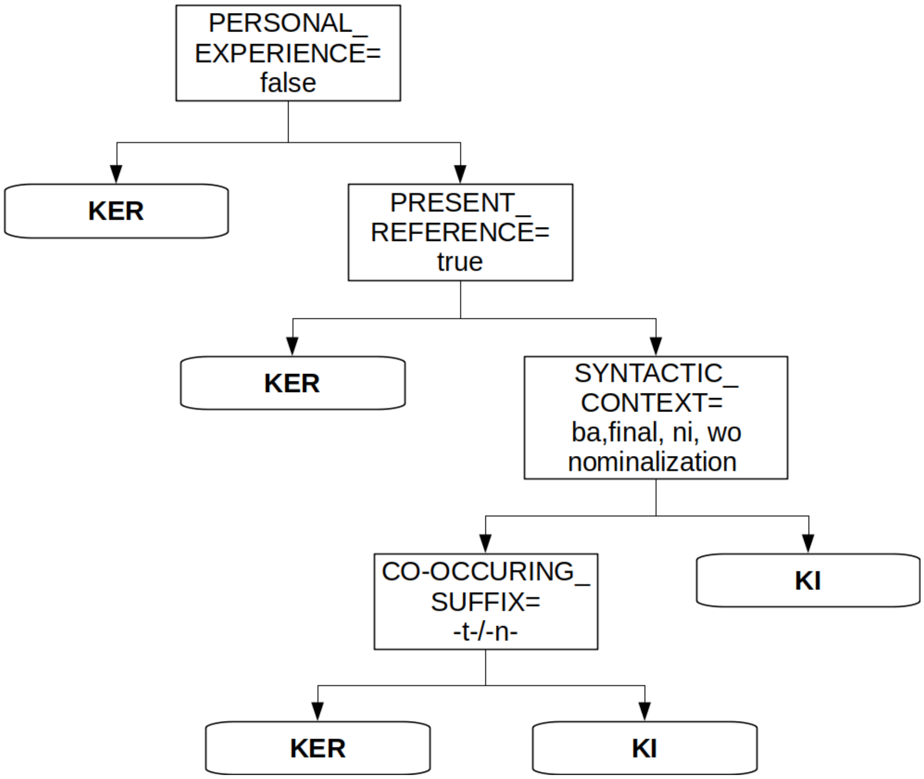

Figure 4.2: Decision tree showing variable interactions; right hand choice when condition is true, left hand choice when false

One thing that immediately stands out is the fact that PERSONAL EXPERIENCE is the root node, whereas the importance estimates in Section 4.1.1 indicate that SYNTACTIC CONTEXT is a more useful predictor overall. Although I have been unable to rigorously confirm it, I believe that this discrepancy is an artefact of the random forest. Due to the multiple levels of depth of SYNTACTIC CONTEXT, it might have been included in more of the classifiers making up the "forest" than the binary PERSONAL EXPERIENCE, and thereby received an importance score out of proportion to its position in the decision tree.

4.2 Analysis

The results in Sections 4.1.1 and 4.1.2 clearly indicate that the major determining factor in the choice between *-ki* or *-ker-* is the variable PERSONAL EXPERIENCE, which codes whether the speaker experienced the event he or she is describing. There is a strong correlation between direct perception of an event and the use *-ki* when retelling it. As such, this provides support for the evidentiality hypothesis as formulated in Takeuchi (1987), Shinzato (1991) and others, whereby *-ki* is an eyewitness past and *-ker-* is an indirect past. Consider:

The inclusion of PRESENT REFERENCE as a high-level node on the right-hand side of the decision tree in Figure 4.2 is also expected, as this variable represents my attempt at operationalizing mirativity. The prototypical data point of this sort tends to look like this:

(19) *ube kaguyafime konomasi-gar-i-tamaf-u-ni koso ar-i-ker-e*
well Kaguyahime like-VERB-0-HON-ATTR-DAT FOC be-0-*ker*-EXCL
"Well, it seems Kaguyahime really likes me!" (TM)

where the predicate is a copula (*-ni ... ar-* in this case) and the whole sentence has connotations of shock or surprise.

Both Shinzato (1991) and Vovin (2003) explicitly address mirativity as a potential contributing factor, although without employing that particular term; and as the reviews in Sandness (1999) and Watanabe (2008) reveal, "exclamative" or "surprise" have long been ascribed to *-ker-* in the Japanese scholarly tradition. Furthermore, the close association between grammaticalized evidentiality and mirativity is a well-known typological fact (DeLancey 1997). The reason for the low importance score of PRESENT REFERENCE in Table 4.3 is likely an artefact of statistics, as only a small minority of data points in my data set assign the value *true* to it.

SYNTACTIC CONTEXT and CO-OCCURRING SUFFIX interact in an interesting way in the rightmost subtree (located under PRESENT REFERENCE=*false*). For verbs in certain syntactic contexts that also bear either of the aspect-signifying suffixes *-t-* or *-n-*, *-ker-* appears even though the larger discourse context would seem to require *-ki*. Data points that conform to this characterization are generally of the type

(20) *ima-fa kafer-ube-ki-ni* ***nar-i-n-i-ker-eba***
now-TOP return-DEB-ATTR-DAT become-0-PFV-0-*ker*-DS
"Now the time has come that I have to go home, so ..." (TM)

or

(21) *isasaka uti-wasure-te* ***ne-ir-i-n-i-ker-u-ni***
somewhat PREF-forget-SS sleep-enter-0-PFV-0-*ker*-ATTR-DAT
"(Waiting for a person to come late at night) I drift off, and then (when I wake up it is already noon)" (MS)

where the verb participates in a clause chain. The impression one gets from reading the entire sentence is that the clause bearing the *-ker-* expresses a circumstance outside the speaker's control, or which is only known retrospectively. I have not coded these examples as PRESENT REFERENCE as their positions within the narrative do not admit such a classification by my criteria. Nevertheless, it is possible that the appearance of *-ker-* is motivated by mirativity. On the other hand, the decision tree algorithm picked up the presence of *-t-* or *-n-* as a distinguishing trait.

To summarize, the statistical modeling reveals that PERSONAL EXPERIENCE and PRESENT REFERENCE are the chief factors underlying the distribution of *-ki* and *-ker-*. Aspect may or may not also be causally linked to the choice of suffix, but if there is a connection, it only holds in circumscribed contexts (i.e. within clause chains). The extent to which COOCCURRING SUFFIX is a useful proxy for a aspect is also debatate (cf. 3.4.1).

4.3 Discussion

The results support the evidentiality hypothesis forwarded in one form or another by Takeuchi (1987), Shinzato (1991) and others, including earlier generations of Japanese scholars whose works I unfortunately was not able to consult (cf. the summaries of previous work in Sandness 1999 and Watanabe 2008).

My analysis relies on a fairly small set of abut 200 data points, in marked contrast to many other quantitative linguistic studies, such as those by Gries (2010), Arppe (2008), and Janda and Lyashevskaya (2011), which all use thousands of samples. Of course, they employed readily available corpuses assembled by teams of linguists over several years, while I only had about two months of data gathering in total. Another issue is that EMJ, being an ancient language, is poorly served in terms of research tools. Tokens of *-ki* and *-ker-* had to be manually collected from texts and annotated, which was tedious and time-consuming in th extreme. Computerized annotation is an attractive alternative, but ultimately infeasible due to the lack of computational resources geared towards EMJ.

Aside from such practical matters, a few larger points regarding methodology deserve mention. As Section 3.4 explains, problem representation is a major issue in machine learning. My choice of variables to include in the model reflects the primary literature on EMJ, and as such it inherits the biases and preconceptions of earlier scholars. Using earlier research as a starting point is reasonable, although it may well be that some unknown factor contributes more to the distribution of *-ki* or *-ker-* than any of the commonly proposed ones. Poplack's (1992) article on the statistics of the French subjunctive illustrates this possibility by showing that none of the traditional "rules" for the subjunctive accurately captures its behavior.

Additionally, operationalization proved to be something of a stumbling block. While I have tried to rely mainly on formal criteria for assigning variable values, some variables depend on an understanding of the larger context in which the sample occurs. PRESENT REFERENCE, BOUNDEDNESS and a few others are suspect in this regard. My interpretation of a passage to a large extent reflects the mainstream opinions of previous scholars, especially seeing as how I turned to translations into contemporary Japanese whenever the original text proved hard to parse. Circular reasoning thus becomes a danger: do I assume that a sentence with *-ki* is PERSONAL EXPERIENCE=*true* because the narrative unambiguously portrays it as such, or because earlier scholars have treated it as such due to the presence of *-ki*? Having multiple human annotators and controlling for inter-annotator disagreement might ameliorate the risk somewhat.

To sum up, hand-coding each sample made it all but impossible to amass a large amount of data, and concerns over operationalization forced me to limit the number of variables included in the model. Despite the overall negative tone of this section, the results were meaningful. The low error rate (22.11%) suggests that the model is on the right track, even if not perfect. Furthermore, the evidence for a strong connection between PERSONAL EXPERIENCE and PRESENT REFERENCE is in line with typologically motivated expectations for evidentiality systems.

Chapter 5

Summary

EMJ has two suffixes with apparent past tense semantics, *-ki* and *-ker-* (Takeuchi 1987 and Trater 2012 add a third, *-t-*). Since they never cooccur on the same verb, and since there is some evidence for a close diachronic relationship (cf. Watanabe 2008), most researchers assume that the forms stand in paradigmatic opposition to one another. The nature of the contrast has proven elusive; speculation largely centers on the notions of evidentiality (e.g. Shinzato 1991, Vovin 2002) and aspect (e.g. Sandness 1999 for *-ki*). The difficulty in pinning down the contrast mirrors the controversies surrounding the interpretation of grammatica categories in other ancient languages, like Biblical Greek (cf. Andrason and Locatell 2016) or Sumerian (Woods 2008:1-44).

In the present study, I attempted to understand the semantics of *-ki* and *-ker-* through quantitative means. Specifically, I sought to identify the factors that are correlated with the use of one of the forms rather than the other in a given context. Traditionally, corpus research on linguistic choice involves utilizing statistical methods to determine the degree of influence a contextual variable exercises over the choice under investigation (Arppe 2008), but following Baayen et al (2013) I elected to use machine learning instead.

A bigger issue than analysis method was that of operationalization and data collection. I used a sample of 194 instances of *-ki* and *-ker-* from three different texts. Each data point was coded for a range of linguistic factors that I deemed to be likely to exert some influence on the "past tense" choice. As per Arppe (2008), I strove to capture as many levels of linguistic analysis as possible: morphological, syntactic, semantic and pragmatic.

The final analysis was executed by means of the random forest algorithm (Ho 1995), in conjunction with decision tree induction to facilitate visualization. The results revealetd that two variables in particular, namely PERSONAL EXPERIENCE and PRESENT REFERENCE, are highly correlated with the outcome of the choice between *-ki* and *-ker-*. Both deal with the speaker or narrator's epistemic stance towards the story being recounted. In particular, PERSONAL EXPERIENCE=*true* is strongly associated with the use of *-ki* as past tense marker, which suggests that *-ki* is a "witnessed past" and *-ker-* is an "unwitnessed past," in agreement with the evidentiality hypothesis (e.g. Takeuchi 1987, Shinzato 1991 and others). Additionally, PRESENT REFERENCE=*true* correlates with -*ker*- even in first-person accounts. Such usages seem to have connotations of surprise, as noted by Shinzato (1991), Sandness (1999) and Vovin (2002) among others.

There was also inconclusive evidence to the effect that aspect plays a minor determining role under particular syntactic and morphological. To be exact, use of -*ker*- is correlated with the presence of the aspect markers -*t*- and -*n*- to the extent of appearing even in contexts where -*ki* would otherwise be expected. As I argue in Section 4.2, this primarily occurs in clause-chaining environments. Unfortunately, I did not have enough data to draw any conclusions about why this may be.

Bibliography

[1] Aikhenvald, A. 2004. *Evidentiality.* Oxford University Press, Oxford.

[2] Akiba, K. 1978. *A Historical Study of Old Japanese Syntax.* PhD thesis, University of Los Angeles, California.

[3] Aksu-Koç, A. A. and Slobin, D. I. 1986. "A Psychological Account of the Development and Use of Evidentials in Turkish." In Chafe, W. and Nichols, J. (eds), *Evidentiality: The Linguistic Coding of Epistemology*, 159-167.

[4] Andrason, A. 2016. "Grammaticalization paths and chaos - Determinism and unpredictability of the semantic development of verbal constructions (Part 2 - Chaos in Linguistics)." *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 133:319-335.

[5] Andrason, A. and Locatell, C. 2016. "The Perfect Wave: A Cognitive Approach to the Greek Verbal System." *Biblical and Ancient Greek Linguistics* 5:7-121.

[6] Anonymous. 2001. *Taketorimonogatari (Zen) (Kadokawa Sofia Bunko: Beginaazu Kurashikkusu.* With translation into modern Japanese by Takeda, T. Tokyo: Kadokawa Shoten.

[7] Arppe, A. 2008. *Univariate, bivariate, and multivariate methods in corpus-based lexicography: A study of synonymy.* PhD thesis, Department of General Linguistics, University of Helsinki.

[8] Baayen, H., Janda, L. A., Nesset, T., Endresen, A. and Makarova, A. 2013. "Making choices in Russian: Pros and cons of statistical methods for rival forms." *Russian Linguistics* 37:253.

[9] Bentley, J. 2001. *A descriptive grammar of early old Japanese prose.* Brill, Leiden; Boston; Köln.

[10] Bentley, J. 2012. "Old Japanese." In Tranter (ed.), *The Languages of Japan and Korea*, pp. 212-245.

[11] Breiman, L. 2001. "Random forests." *Machine Learning* 45:5-32.

[12] Curnow, T. 2001. “Evidentiality and Me: The Interaction of Evidentials and Person.” *Proceedings of the 2001 Conference of the Australian Linguistic Society*

[13] DeLancey, S. 1997. “Mirativity: The grammatical marking of unexpected information.” *Linguistic Typology* 1:33-52.

[14] Denil, M., Matheson, D. and de Freitas, N. 2014. “Narrowing the Gap: Random Forests in Theory and In Practice.” *Proceedings of the 31th International Conference on Machine Learning*, Beijing, China.

[15] Divjak, D. and Gries, S. Th. 2006. “Ways of trying in Russian: clustering behavioral profiles.” *Corpus Linguistics and Linguistic Theory* 2:23-60.

[16] Frellesvig, B. 2010. *A History of the Japanese Language.* Cambridge University Press, Cambridge.

[17] Gries, S. Th. 2010. “Behavioral profiles. A fine-grained and quantitative approach in corpus-based lexical semantics.” *The Mental Lexicon* 5:3, pp. 323-346.

[18] Gries, S. Th. 2014. *Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us.* In Taavitsainen, I., Kytö, M., Claridge, C., and Smith, J. (eds.), *Developments in English: expanding electronic evidence*; pages 29-47.

[19] Ho, T. K. 1995. “Random Decision Forests.” *Proceedings of the 3rd Conference on Document Analysis and Recognition*, Montreal, QC, pages 278-282.

[20] Hard, A. 2019. “jpdata.” Github commit. Available online: https://github.com/aehard/jpdata

[21] Janda, L. A. and Lyashevskaya, O. 2011. “Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian.” *Cognitive Linguistics* 4:719-763.

[22] Jansegers, M. and Gries, S. Th. 2017. “Towards a dynamic behavioral profile: A diachronic study of polysemous *sentir* in Spanish.” *Corpus Linguistics and Linguistic Theory.*

[23] Manning, C. D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.* Massachusetts Institute of Technology, Massachusetts.

[24] Murasaki Shikibu. 1999. *Genji Monogatari.* Romanization, translation into modern Japanese by anonymous. URL: http://jti.lib.virginia.edu/japanese/genji/index.html.

[25] Lanz, L. A. 2009. “Diachrony of complex predicates in Japanese.” *Rice Working Papers in Linguistics* 1:168.

[26] Liaw, A. and Wiener, M. 2002. “Classification and Regression by randomForest.” *R News* 2:3, 18-22.

[27] Louppe, G., Wehenkel, L., Sutera, A. and Geurts, P. 2013. “Understanding variable importances in Forests of randomized trees.” NIPS 13 Proceedings of the 26th International Conference on Neural Information Processing Systems 1:431-439.

[28] Ohori, T. 1992. *Diachrony in Clause Linkage and Related Issues.* PhD thesis, University of California at Berkeley.

[29] Poplack, S. 1992. “The inherent variability of the French subjunctive.” In Laeufer, C. and Terrell, M. (eds), *Theoretical Analyses in Romance Linguistics.* Amsterdam: Benjamins. 235-263.

[30] Quinlan, J. R. 1986. “Induction of Decision Trees.” *Machine Learning* 1:81-106.

[31] R Core Team. 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL: http://www.R-project.org/

[32] Redman, T. C. 2018. “If Your Data Is Bad, Your Machine Learning Tools Are Useless.” *Harvard Business Review*, https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Accessed 28th Nov. 2018.

[33] Samuel, A. L. 1959. “Some studies in machine learning using the game of checkers.” *IBM Journal of Research and Development* 44:22, p. 206-226.

[34] Sandness, K. 1999. *The Evolution of the Japanese Past and Perfective Suffixes.* University of Michigan Press.

[35] Sei Shonagon. 2001. *Makura no Sōshi (Kadokawa Sofia Bunko: Biginaazu Kurashikkusu).* With translation into modern Japanese by Sakaguchi, Y. Tokyo: Kadokawa Shoten.

[36] Shinzato, R. 1991. “Where do temporality, evidentiality and epistemicity meet? A comparison of Old Japanese *-ki* and *-keri* with Turkish *-di* and *-miş*.” *Gengo Kenkyuu* 95:25.

[37] Takeuchi, L. 1987. *A Study of Classical Japanese Tense and Aspect.* Akademisk Forlag, Copenhagen.

[38] Tranter, N. 2012. “Classical Japanese.” In Tranter (ed.), *The Languages of Japan and Korea*, Routledge, pp. 212-245.

[39] Tranter, N. and Kizu, M. 2012. “Modern Japanese.” In Tranter (ed.), *The Languages of Japan and Korea*, Routledge, 268-313.

[40] Vendler, Z. 1957. “Verbs and times.” *The Philosophical Review* 66:143-160.

[41] Vovin, A. 2002. *A Reference Grammar of Classical Japanese Prose*. Routledge Curzon, New York.

[42] Vovin, A. 1997. “On the syntactic typology of Old Japanese.” *Journal of East Asian Linguistics* 6:273-290.

[43] Watanabe, K. 2008. *Tense and Aspect in Old Japanese: Synchronic, Diachronic and Typological Perspectives*. PhD thesis, Cornell University.

[44] Woods, C. 2008. *The Grammar of Perspective. The Sumerian Conjugation Prefixes as a System of Voice*. Brill, Leiden.