1

Introduction to Japanese Computational Linguistics

Francis Bond and Timothy Baldwin

The purpose of this chapter is to provide a brief introduction to the Japanese language, and natural language processing (NLP) research on Japanese. For a more complete but accessible description of the Japanese language, we refer the reader to Shibatani (1990), Backhouse (1993), Tsujimura (2006), Yamaguchi (2007), and Iwasaki (2013).

1 A Basic Introduction to the Japanese Language

Japanese is the official language of Japan, and belongs to the Japanese language family (Gordon, Jr., 2005).[1] The first-language speaker population of Japanese is around 120 million, based almost exclusively in Japan.

The official version of Japanese, e.g. used in official settings and by the media, is called *hyōjuNgo* "standard language", but Japanese also has a large number of distinctive regional dialects. Other than lexical distinctions, common features distinguishing Japanese dialects are case markers, discourse connectives and verb endings (Kokuritsu Kokugo Kenkyujyo, 1989–2006).

[1]There are a number of other languages in the Japanese language family of Ryukyuan type, spoken in the islands of Okinawa. Other languages native to Japan are Ainu (an isolated language spoken in northern Japan, and now almost extinct: Shibatani (1990)) and Japanese Sign Language.

Readings in Japanese Natural Language Processing.
Francis Bond, Timothy Baldwin,
Kentaro Inui, Shun Ishizaki,
Hiroshi Nakagawa and Akira Shimazu (eds.).

2 The Sound System

Japanese has a relatively simple sound system, made up of 5 vowel phonemes (/a/,[2] /i/, /u/, /e/ and /o/), 9 unvoiced consonant phonemes (/k/, /s/,[3] /t/,[4] /n/, /h/,[5] /m/, /j/, /ɽ/ and /w/), 4 voiced consonants (/g/, /z/,[6] /d/ [7] and /b/), and one semi-voiced consonant (/p/). These phonemes combine to make up syllables composed as follows: (1) an onset of zero, one or two consonants; (2) one of the five vowels; and (3) optionally a coda, in the form of an optional chroneme (lengthened vowel[8]) and the optional consonant /n/ (with the chroneme preceding /n/ if they both occur). For example, the syllable /koon/ is made up of the onset /k/, vowel /o/, and onset made up of a chroneme (/o/) and final consonant /n/. Double-consonant onsets take the form of any consonant other than /j/ and /w/, combined with /j/, e.g. /gjuunjuu/ (Romanized as *gyūnyū*, and meaning "milk").[9]

3 The Writing System

The Japanese writing system is made up of three separate sets of characters: hiragana, katakana and kanji. Modern Japanese also commonly makes use of Arabic numbers and Latin script (e.g. in company and product names, or in rendering the names of non-Japanese entities).

Hiragana and **katakana** (collectively referred to as **kana**) are isomorphic syllabaries made up of 46 basic characters, made up of: (1) the five standalone vowels (*a* (あ), *i* (い), *u* (う), *e* (え) and *o* (お), in alphabetical order); (2) single-consonant–vowel syllables (e.g. *ka* (か), *ni* (に) or *yo* (よ)); and (3) the single-character nasal sonorant (*N* (ん)). The 46 characters are arranged in a 10×5 grid (with some gaps) called the *gojūoN* "fifty sounds" as presented in Table 1, based on the 5 standalone vowels along with the combination of those vowels with each of 9 character-initial consonants (*k*, *s*, *t*, *n*, *h*, *m*, *y*, *r*, *w*);[10] this grid also forms the basis of the standard alphabetic ordering of Japanese,

[2]For a general introduction to phonetic transcription, see Clark et al. (2007), and for an introduction to Japanese phonology, see Vance (1987).

[3]Pronounced [ɕ] when it precedes /i/.

[4]Pronounced [tɕ] when it precedes /i/ and [ts] when it precedes /u/.

[5]Pronounced [ɸ] when it precedes /u/.

[6]Pronounced [dʑ] when it precedes /i/.

[7]Pronounced [dʑ] when it precedes /i/ and [z] when it precedes /u/.

[8]Commonly indicated in transliterated Japanese with ¯ or ˆ, although this is often lost in their English renderings: for example, Tokyo is actually *Tōkyō* (both *o*'s are long vowels), and judo is actually *jūdō*.

[9]In modern Japanese, only /a/, /u/ and /o/ combine with double-consonant offsets.

[10]Of these, five consonant–vowel combinations are not included in the sound system of modern Japanese, and do not have corresponding kana, namely *wi*, *wu*, *we*, *yi* and *ye*.

Table 1: Basic hiragana and their corresponding romanizations, in orthographic order (top-down, left-to-right across the two tables)

		Onset					
		—	/k/	/s/	/t/	/n/	/h/
Vowel	/a/	あ *a*	か *ka*	さ *sa*	た *ta*	な *na*	は *ha*
	/i/	い *i*	き *ki*	し *shi*	ち *chi*	に *ni*	ひ *hi*
	/u/	う *u*	く *ku*	す *su*	つ *tsu*	ぬ *nu*	ふ *fu*
	/e/	え *e*	け *ke*	せ *se*	て *te*	ね *ne*	へ *he*
	/o/	お *o*	こ *ko*	そ *so*	と *to*	の *no*	ほ *ho*

		Onset				
		/m/	/j/	/ɾ/	/w/	/n/
Vowel	/a/	ま *ma*	や *ya*	ら *ra*	わ *wa*	
	/i/	み *mi*		り *ri*		
	/u/	む *mu*	ゆ *yu*	る *ru*		
	/e/	め *me*		れ *re*		
	/o/	も *mo*	よ *yo*	ろ *ro*	を *wo*	
	—					ん *N*

Table 2: Voiced and semi-voiced hiragana, and their corresponding romanizations

		Onset				
		/g/	/z/	/d/	/b/	/p/
Vowel	/a/	が *ga*	ざ *za*	だ *da*	ば *ba*	ぱ *pa*
	/i/	ぎ *gi*	じ *ji*	ぢ *ji*	び *bi*	ぴ *pi*
	/u/	ぐ *gu*	ず *zu*	づ *zu*	ぶ *bu*	ぷ *pu*
	/e/	げ *ge*	ぜ *ze*	で *de*	べ *be*	ぺ *pe*
	/o/	ご *go*	ぞ *zo*	ど *do*	ぼ *bo*	ぽ *po*

Table 3: Double-consonant onset compound hiragana (both unvoiced and (semi-)voiced) and their corresponding romanizations

		Onset					
		/kj/	/sj/	/tj/	/nj/	/hj/	/mj/
Vowel	/a/	きゃ *kya*	しゃ *sha*	ちゃ *cha*	にゃ *nya*	ひゃ *hya*	みゃ *mya*
	/u/	きゅ *kyu*	しゅ *shu*	ちゅ *chu*	にゅ *nyu*	ひゅ *hyu*	みゅ *myu*
	/o/	きょ *kyo*	しょ *sho*	ちょ *cho*	にょ *nyo*	ひょ *hyo*	みょ *myo*

		Onset				
		/ɾj/	/gj/	/zj/	/bj/	/pj/
Vowel	/a/	りゃ *rya*	ぎゃ *gya*	じゃ *ja*	びゃ *bya*	ぴゃ *pya*
	/u/	りゅ *ryu*	ぎゅ *gyu*	じゅ *ju*	びゅ *byu*	ぴゅ *pyu*
	/o/	りょ *ryo*	ぎょ *gyo*	じょ *jo*	びょ *byo*	ぴょ *pyo*

working down each column, left-to-right across the columns from *a* to *N*. Additional syllables are constructed by voicing or semi-voicing the consonant by attaching a *dakuteN* (゛) or *haNdakuteN* (゜), respectively, to the top-right of the character (e.g. producing *ga* (が) as the voiced variant of *ka* (か), and *pe* (ぺ) as the semi-voiced variant of *he* (へ)), as presented in Table 2. Two-consonant onsets are lexicalized by appending *ya* (や), *yu* (ゆ) or *yo* (よ) in smaller font to the character corresponding to the first consonant combined with *i*, as outlined in Table 3; for example *kyu* is formed by combining *ki* (き) with *yu* (ゆ), i.e. きゆ. These compound characters can optionally be (semi-)voiced by appending a *dakuteN* or *haNdakuteN* to the first character (e.g. ぎゆ for *gyu*).

While hiragana and katakana are termed **syllabaries**, the basic unit is a technically a **mora**: a sound unit of roughly constant length. A single syllable with a long vowel sound is made up of two morae: a standalone vowel is appended to the base character cluster. For example, *kyū* is formed by appending *u* (う) to *kyu* (きゆ), i.e. きゅう.[11] Thus a single syllable can, in practice, be made up of multiple kana characters: by vowel lengthening, combing two characters for a complex onset and/or adding a final N.

The third character system is **kanji**, and is **logogrammatic** in nature, i.e. individual characters represent single morphemes, such as *ichi* (一) "one" or *dō* (動) "motion, change". The standard estimate for the number of kanji characters that are commonly used in writing Japanese is 2,136, based on the set of Joyo Kanji stipulated by the Japanese Ministry of Education, Culture, Sports, Science and Technology to be taught in Japanese primary and high schools. Thousands more are used in place names, person names and historical texts.

A single kanji character generally has at least one **on**-reading which is loosely derived from its Chinese pronunciation at the time of borrowing,[12] and at least one native Japanese **kun**-reading where a Japanese word which pre-existed the orthographic borrowing was mapped onto a kanji character based on rough semantic correspondence. For example, 動 has a unique on-reading of *dō*, and a unique kun-reading of *ugo(ku/kasu)*;[13] in both cases, its basic meaning is "motion, change".

[11] In most words with a long *ō* vowel, the vowel lengthening is indicated with the character *u* (う) rather than *o* (お) (e.g. *kō* (こう), but note *ōkī* (おおきい) "large"). In katakana, the character ー is often used to lengthen the vowel of the preceding character (e.g. *ko* (コ) vs. *kō* (コー)).

[12] Indeed, many kanji still have corresponding hanzi in traditional Chinese, although there are also a few kanji which were devised in Japan and are unique to Japanese, such as *hatake* (畑) "field" and *tōge* (峠) "mountain pass".

[13] The reading of 動 itself is *ugo*, and it combines with a kana-based conjuga-

We have characterized kanji as logograms, and indeed many kanji can occur as single-character morphemes in text, generally pronounced using their kun-reading (e.g. *kokoro* (心) "heart, spirit") and often with okurigana (especially for verbs and adjectives, e.g. *ugoita* (動いた) "moved (intrans.)" or *omoi* (重い) "heavy"). More commonly, however, kanji combine with other kanji to form multi-kanji morphemes such as *shiNkyō* (心境) "mental state") or *jūshiN* (重心) "centre of gravity, centroid". Two-kanji morphemes, in particular, are very common in Japanese. The readings of multi-kanji morphemes are almost always formed compositionally from the readings of the component characters (Yencken and Baldwin, 2005), generally comprising all on- or all kun-readings. Composition of the readings is often accompanied by **sequential voicing** or **gemination**. In sequential voicing (known as *reNdaku* in Japanese), a kanji with trailing consonant /n/ is immediately followed by a kanji with a "voiceable" leading consonant (i.e. /k/, /s/, /t/ or /h/), and the leading consonant is voiced (Yamaguchi, 2007). For example, *kaN* (肝) "liver" + *shiN* (心) "heart, spirit" = *kaNjiN* (肝心) "essential". Note that the sequential voicing is not marked on the kanji in any way (although it would, of course, be reflected in the kana rendering of the word). Gemination can be thought of as the equivalent process for consonants, whereby the final mora (usually ending in the vowel /u/) of the leading kanji is dropped, to be replaced by the leading consonant of the trailing kanji (Vance, 1987); for example, *ketsu* (決) "decide" + *shiN* (心) "heart, spirit" = *kesshiN* (決心) "determination, resolution".[14] While rare, there are also instances of multi-kanji morphemes with non-compositional readings, such as *dai* (台) "table, support" + *shi* (詞) "words, lyrics" = *serifu* (台詞) "speech, lines". More common are multi-kanji words which are *semantically* non-compositional, as seen with the examples *kaNjiN* (肝心) "essential" and *serifu* (台詞) "speech, lines" above.

In standard Japanese text, hiragana is primarily used for function words, auxiliary words, manner words (e.g. onomatapeic expressions) and for transcribing rare kanji. Katakana is standardly used for transliterations of foreign words — of which there are many in Japanese (e.g. *supōtsu* (スポーツ) "sport" or *heddohōN* (ヘッドホーン) "head-

tional suffix (**okurigana**) derived from *ku* or *kasu* (corresponding to intransitive and transitive verb usages, respectively), e.g. *ugoita* (動いた) "moved (intrans.)" or *ugokashiteiru* (動かしている) "is moving (trans.)".

[14] Gemination is marked in the kana rendering of the word by っ (named *sokuoN*, but with no standalone pronunciation). To repeat our example of gemination in hiragana, therefore: *ketsu* (けつ) "decide" + *shiN* (しん) "heart, spirit" = *kesshiN* (けっしん) "determination, resolution".

phones") — and scientific names of plants and animals, and sometimes for emphasis, much as italics are used in English. Kanji is reserved for the stems of content words. As such, the three character systems are interspersed in standard Japanese writing, e.g.:[15]

(1) コアラが　寝　た
koara *ga* *ne* *ta*
koala NOM sleep PAST
"The koala slept"

where the first morpheme (*koara* (コアラ)) is in katakana due to it being a transliterated borrowing, the second and fourth morphemes (the case particle *ga* (が) and tense marker *ta* (た)) are in hiragana due to them being function words, and the third morpheme *ne* (寝) is in kanji.

4 Morphosyntax

Japanese is a verb-final language, which marks arguments for grammatical/semantic role with postpositional **case markers** (a.k.a. **case particles**, **postpositions** , or simply **particles**). For example, in Example (1), the verb *tabeta* can be seen to occur at the end of the clause, the subject *koara* is marked with the nominative case marker *ga*, the object *happa* is marked with the accusative case marker *o*,[16] and the adverb *yukkuri* is marked with the manner case marker *to*:

(2) コアラが　葉っぱを　ゆっくりと　食べた
koara *ga* *happa* *o* *yukkuri* *to* *tabe* *ta*
koala NOM leaf ACC slowly MAN eat PAST
"The koala slowly ate a leaf"

Other than in colloquial spoken Japanese or marked styles such as headlines, all complements and most adjuncts are marked with a case marker,[17] making it possible to **scramble** the order of the case-marked constituents and still recover the argument structure of the clause. As

[15]For details of the notations used in interlinear-glossed text examples in this book, see the table at the start of the book (page viii).

[16]The observant reader will recall that in Section 3, we listed the hiragana character used to mark the object (を) as being pronounced *wo*. This character is used almost exclusively as a case marker, in which instance it is pronounced *o*.

[17]The *to* case marker on the adverbial *yukkuri* is optional in Example (2), but without it, scrambled word orders where the adverb is not adjacent to the verb are ungrammatical, or at least unnatural. Temporal adjuncts (e.g. *kyō* "today") are also typically not case marked.

such, all of the following are grammatical Japanese and almost identical in meaning to the original in Example (2) (modulo the effects of information structure; see Section 5):

(3) a. 葉っぱを コアラが ゆっくりと 食べた
happa o koara ga yukkuri to tabe ta
leaf ACC koala NOM slowly MAN eat PAST

b. ゆっくりと 葉っぱを コアラが 食べた
yukkuri to happa o koara ga tabe ta
slowly MAN leaf ACC koala NOM eat PAST

c. ゆっくりと コアラが 葉っぱを 食べた
yukkuri to koara ga happa o tabe ta
slowly MAN koala NOM leaf ACC eat PAST

On this basis, Japanese is often described as a **free word order language**. Note, however, that word order scrambling is subject to a number of constraints, including leaving the verb at the end of the clause,[18] moving constituents in their entirety (including the case particle), and moving constituents only within the boundaries of the clause they are contained in. For example, the following are not grammatical Japanese (due to the main verb not being clause-final in (a) and a constituent being separated from its case particle in (b)):

(4) a. ∗ 食べた コアラが 葉っぱを ゆっくりと
tabe ta koara ga happa o yukkuri to
eat PAST koala NOM leaf ACC slowly MAN

b. ∗ 葉っぱコアラが を ゆっくりと 食べた
happa koara ga o yukkuri to tabe ta
leaf koala NOM ACC slowly MAN eat PAST

Also note that there will tend to be a default order for a given set of constituents and verb. As a broad generalization, where a constituent of the indicated type is present, the default constituent order tends to be:[19]

1. Topic (e.g. *saikiN no yononaka-wa* "the modern world") — see Section 5

[18] Other than in informal speech, where case-marked arguments can be uttered after the main verb in speech repairs or to post-hoc resolve zero anaphora-based ambiguity in the utterance.

[19] The example constituents make up the sentence *saikiN no yononaka-wa dekiru hito-ga deki nai hito-ni shigoto-o jikkuri oshie naku nat ta yō da* "In the modern world, capable people no longer seem to (have the time) to teach their job to those less capable than them".

2. Subject (e.g. *dekiru hito-ga* "capable people")
3. Indirect object (e.g. *deki nai hito-ni* "incapable people")
4. Direct object (e.g. *shigoto-o* "work")
5. Manner (e.g. *jikkuri* "patiently, carefully")
6. Predicate (e.g. *oshie naku nat ta yō da* "seem to no longer teach")

There is a weak constituent order preference for (non-topicalized) temporal and locative constituents to occur at the start or end of the clause (just before the main verb).

Similarly to Chinese and Thai, written Japanese is non-segmented, i.e. morpheme boundaries are not overtly marked. As such, the native rendering of Example (2) is コアラが葉っぱをゆっくり食べた, with no indication of where morphemes start and end. Because of the lack of word segmentation, the notion of **word** is somewhat ill-defined in Japanese. For example, *kaikeibuchō* (会計部長) "accounting department head" is made up of the three morphemes *kaikei* (会計) "accounting", *bu* (部) "department" and *chō* (長) "head". It is possible to analyse the three-morpheme compound as either left-branching (i.e. *((kaikei bu) chō)* "((accounting department) head)") or right-branching (i.e. *(kaikei (bu chō))* "(accounting (department head))"), with each suggesting a different "word" analysis. The semantics of these two analyses is largely indistinguishable, however. Ultimately, therefore, the internal structure of the compound is underspecified, and there is no easy answer to the question of what "words" it is made up of.

The predominant word classes in Japanese are as follows:

nouns (N) (e.g. *koara* "koala" and *happa* "leaf") — non-conjugating; no marking for number (e.g. singular vs. plural)[20] or grammatical gender or definiteness (Bond, 2005); highly productive right-headed noun compounding via simple concatenation (e.g. *kikai* "machine" + *hoNyaku* "translation" + *kyōkai* "association" = *kikai hoNyaku kyōkai* "machine translation association") or linking with the *no* case marker (e.g. *kaisha* "company" + *hito* "person" = *kaisha no hito* "company person": Tanaka and Baldwin (2003))

verbs (V) (e.g. *ugo(ku)* "move (intrans.)" and *kie(ru)* "extinguish, disappear") — conjugating, largely via regular conjugation classes as indicated by the suffix in parentheses (i.e. *ugo(ku)* "move (intrans.)", *ka(ku)* "write" and *ugome(ku)* "wriggle" all conjugate identically); past vs. nonpast tense; passivization etc. marked

[20] Although there are (optional) suffixes such as *-tachi* for human-referent nouns which indicate a group (e.g. *hito-tachi* "group of students".

synthetically with auxiliary verbs (see below); highly productive verb–verb compounding (Uchiyama et al., 2005; Nishiyama, 2008; Breen and Baldwin, 2009); no marking of agreement with the subject or other arguments[21]

verbal nouns (NS)[22] (e.g. *kesshiN* "determination, resolution" and *iNshoku* "eat and drink") — when used as a noun, shares all of the properties for nouns listed above; can also be used as a denominal verb primarily in combination with the light verb *suru* "do", optionally with accusative case marking (i.e. as either *kesshiN suru* or *kesshiN o suru*, both meaning "decide, resolve": Miyamoto (1999)), in which case the light verb construction shares the properties of verbs listed above

adjectives (A) (e.g. *oishī* "tasty" and *nagai* "long") — can be used attributively (as a pre-modifier, e.g. *oishī gohaN* "tasty food") and predicatively (e.g. *gohaN ga oishī* "food is tasty") usages; predicative adjectives take case-marked arguments similarly to verbs, but with a restricted set of case markers and nominative marking for the object (e.g. *tōkyō ga gohaN ga oishī* "Tokyo food is tasty"); conjugate for tense (past vs. nonpast); adverb form derivable by conjugation of final *i* to *ku* (e.g. *nagaku matsu* "wait a long (time)")

adjectival noun (AN)[23] (e.g. *gaNko* "stubborn" and *jōbu* "strong, robust") — like adjectives, can be used attributively (as a pre-modifier, with the *na* auxiliary, e.g. *gaNko na seikaku* "stubborn Personality") and predicatively (e.g. *seikaku ga gaNko* "personality is stubborn"); argument-taking properties largely the same as adjectives; no conjugation for tense (tense marking is via the insertion of a copula verb); adverb form derivable with the *ni* case marker (e.g. *gaNko ni matsu* "wait stubbornly")

adverbs (RB) (e.g. *sugu* "immediately" and *kanari* "fairly") — directly premodify adjectives and adjectival nouns (e.g. *kanari nagai* "fairly long" and *kanari gaNko* "fairly stubborn"); modification of verbs without case marking (e.g. *tokidoki kieru* "sometimes disappear", optionally with *to* case marking (e.g. *yukkuri to kieru* "gradually disappear"), or (optionally) with *ni* case marking (e.g. *sugu ni kieru* "immediately disappear"), depending on the adverb; large numbers of onomatopoeic adverbs (e.g. *mekimeki* "visibly" and *suisui* "gracefully, smoothly")

[21] Although there are agreement-like effects with certain adverbs (e.g. *chittomo* "not at all") or postpositional modifiers (e.g. *shika* "only") requiring the verb to have positive or negative polarity.

pronouns (Pro)[24] (e.g. *watashi* "I" and *sore* "that") — no marking for grammatical case; implicitly singular number (e.g. *watashi* can only refer to the singular first person; to refer to the plural first person, a group-marking suffix such as *tachi* must be used: Bond (2005)), other than for overtly plural pronouns (e.g. *wareware* "we"); relatively free pre-modification possible (e.g. *odoroita kare* "lit: the surprised he"); heavy politeness marking (see Section 5); person-referent pronouns are much more common than object-reference pronouns (where zero anaphors are more common; see Section 5)

classifiers (CL) (e.g. *dai* "machines" and *hoN* "long thin objects") — when enumerating most objects in Japanese, numerals must combine with a classifier specifying the semantic type of the object (Downing, 1996); number–classifier clusters can pre-modify nouns, usually with the *no* case marker (e.g. *2 dai* "2 machines" + *puriNtā* "printer" = *2 dai no puriNtā* "2 printers"; c.f. ∗*2 dai puriNtā*) or post-modify (case-marked) nouns (e.g. *puriNtā o kau* "buy a printer" + *2 dai* "2 machines" = *2 dai no puriNtā o kau* "buy 2 printers" or *puriNtā o 2 dai kau* "buy 2 printers"); dozens of classifiers in common use, and strong sortal constraints on classifier compatibility for most referents

case particles (P) (e.g. *ga* "NOM" and *kara* "from") — post-modify noun phrases, and some adjectival and adverbial phrases; some case particles act most like markers of grammatical role (e.g. *o* which mostly marks objects of verbs), while others are markers of the semantics of adjuncts (e.g. *made* mostly marks spatio-temporal destination NPs)

clause-final particles (PF) (e.g. *ka* "Q" and *no* "NML") — post-modify clauses to indicate clause type (e.g. the interrogative *koara ga tabe ta ka* "did the koala eat?" is formed from the declarative *koara ga tabe ta* "the koala ate" by the addition of the clause-final particle *ka*) or nominalization (e.g. *koara ga tabe ru* "a koala eats" + *mi ta* "(I) saw" = *koara ga tabe ru no o mita* "I saw a koala eating")

auxiliary verbs (VA) (e.g. *(r)are* "PASS" and *na(i)* "NEG") — post-modify verb stems according to the conjugation class of the stem, to indicate passivization (e.g. *ka kare ta* "written"), negation (e.g. *ugo ka na i* "doesn't move"), potential (e.g. *ugo ke ta* "could move") and other verb modality/aspect.

adnominal modifier (RT) (e.g. *kono* "this" and *aru* "certain") — pre-modifiers to nouns, which can be used to mark definiteness

or specificity, or locate referents relative to the speaker/addressee (known as *reNtaishi* in Japanese)

Example (5) is an example of a sentence which includes instances of all these word classes (with the parts of speech marked in the third gloss line based on the acronyms listed above for each word class):

(5)

2	匹	の	コアラが		その	青	い		葉っぱを	
2	*hiki*	*no*	*koara*	*ga*	*sono*	*ao*	*i*		*happa*	*o*
1	animal	GEN	koala	NOM	those	green	NONPAST		leaf	ACC
N	CL	P	N	P	RT	A			N	P
ゆっくりと		食べて		頑固		に	私	を	無視	し
yukkuri	*to*	*tabe*	*te*	*gaNko*		*ni*	*watashi*	*o*	*mushi*	*shi*
slowly	MAN	eat	TE	stubborn		DAT	me	ACC	ignore	do
RB	P	V		AN		P	Pro	P	NS	VA
続け	た	よ								
tsuduke	*ta*	*yo*								
continue	PAST	EXCL								
VA		PF								

"Two koalas slowly ate those green leaves and obstinately continued to ignore me!"

As with other languages, Japanese is rich in multiword expressions (MWEs: Sag et al. (2002); Baldwin and Kim (2009)), including noun–noun compounds, verb–verb compounds and light verb constructions (as mentioned above in this section), and also multiword case particles (e.g. *ni tsuite* "concerning": Baldwin and Bond (2002)), four-character idiomatic compounds borrowed from Chinese, verbal idioms (e.g. *ude-o age(ru)* "raise one's skill level": Hashimoto and Kawahara (2008); Shudo et al. (2011); Fothergill and Baldwin (2011, 2012)), and lexical borrowings from languages such as English which have been transliterated wholesale (e.g. *sekusharu-harasumeNto* "sexual harassment") or constructed from other lexical borrowings in ways which deviate from the source language (termed *wasei-eigo* "Japan-made English", e.g. *waN-patāN* "repetitive, monotone (lit: one pattern)": Breen et al. (2012)).

Accounts of Japanese phonosyntax are often founded on the notion of **bunsetsu**, which are made up of a single or compound content word, and any right-attached function words and case particles (and left-attached politeness markers). Bunsetsu relate closely to **chunks** in English and other languages in that they are sub-phrasal and right-headed (or at least the rightmost content word in a bunsetsu is the semantic head), and were originally developed in the context of anal-

ysis of how Japanese is read and spoken (starter readers in Japanese often incorporate whitespaced-based bunsetsu boundaries for readability purposes). Consider Example (6), for example:

(6) 動物　保護　区　の　コアラが　青　い
dōbutsu hogo ku no koara ga ao i
animal sanctuary zone GEN koala NOM green NONPAST
葉っぱを　ゆっくりと　食べた
happa o yukkuri to tabe ta
leaf ACC slowly MAN eat PAST
"The koala at the animal sanctuary slowly ate a green leaf"

The bunsetsu structure of Example (6) is as follows (with bunsetsu boundaries indicated with spaces, and intra-bunsetsu morpheme boundaries indicated with hyphens):

(7) 動物-保護-区-の　コアラ-が　青-い
dōbutsu-hogo-ku-no koara-ga ao-i
animal-sanctuary-zone-GEN koala-NOM green-NONPAST
葉っぱ-をゆっくり-と食べ-た
happa-o yukkuri-to tabe-ta
leaf-ACC slowly-MAN eat-PAST
"The koala at the animal sanctuary slowly ate a green leaf"

Of note are: (1) the noun compound in the first bunsetsu combining into a single bunsetsu (*dōbutsu-hogo-ku-no*); (2) the fact that the NP subject is split up into two bunsetsu because of the genitive case marker (*no*), similarly to what happens to NPs containing possessives in English (e.g. *[the koala] ['s appetite]*, except that *no* attaches to the *preceding* bunsetsu; and (3) attributive adjectives (*aoi*) form their own bunsetsu, unlike attributive adjectives in English which are incorporated into noun chunks (e.g. *[the green leaves]*).

When bunsetsu are used as the basis of syntactic trees, they are assumed to always modify a bunsetsu to the right of them. Whether bunsetsu are used as the basis of syntactic analysis or not, due to the verb-final nature of the language, Japanese phrase structure trees tend to be heavily left branching.

5 Pragmatics and Sociolinguistics in Japanese

Japanese makes heavy use of **zero anaphora**, in omitting constituents (case marker and all) in contexts where the constituent can be recovered (Kameyama, 1985). For example, if the following sentence were to follow Example (2):

(8) コアラが 水 を 飲ま な かった
koara ga mizu o noma na katta
koala NOM water ACC drink NEG PAST
"The koala didn't drink water"

a more natural realization would be:[25]

(9) 水 を 飲ま な かった
mizu o noma na katta
water ACC drink NEG PAST
"(It) didn't drink water"

where the subject (*koara ga*) has been omitted entirely, on the basis that it can be recovered from the discourse context. There are almost no restrictions on what arguments can be elided. For example, the following is perfectly well-formed Japanese:

(10) 渡し た
watashi ta
hand PAST
"(I) handed (it) (to someone)"

where the subject, direct object and indirect object (and possibly adjuncts such as the time and location of the event) have been elided, but are potentially resolvable from discourse context. If it were in response to Example (11), for example, it could be readily interpretable as "(Yes,) (I) handed (the paper) (to my professor) (at university) (yesterday)".

(11) 昨日 論文 を 大学 で 先生 に 渡し
kinō roNbuN o daigaku de seNsei ni watashi
yesterday paper ACC university LOC professor DAT hand
た か
ta ka
PAST Q
"Did (you) hand the paper to your professor at university yesterday?"

It is also possible to omit constituents without explicit mention of them in the discourse context, where they can be inferred through extralinguistic context or real-world knowledge.

[25]Perhaps a more natural realization again would be to topicalize *mizu*, as we'll discuss later in this section.

As can be seen in the translation of the subject in Example (10), particular argument positions often have strong default referents based on a combination of factors including the governing verb, clause type (e.g. if the clause were interrogative rather than declarative the default subject would be the addressee), and empathy marking (see below in this section). In fact, with first- and second-person referents in particular, these defaults are often so strong that it can be unnatural-sounding in Japanese to realize default-interpretable arguments subject with overt noun phrases such as *watashi ga* or similar.

Japanese also makes heavy use of topicalization, and is often categorized as a **topic–comment language** (Kitagawa, 1982; Shibatani, 1991). The primary means of topicalization is in an argument being promoted to the topic and marked with the topic marker *wa*,[26] whereby the original case marker is either replaced by *wa* (in the case of *ga* and *o*) or *wa* is appended to the original case marker (in the case of other case markers, e.g. *ni* becomes *ni wa*). Topicalized constituents are often (but not always; see Example (13)) moved to the front of the clause. For example:

(12) a. コアラが　葉っぱを　ゆっくりと　食べた
koara ga happa o yukkuri to tabe ta
koala NOM leaf ACC slowly MAN eat PAST
"The koala slowly ate a leaf"

b. 葉っぱは　コアラが　ゆっくりと　食べた
happa wa koara ga yukkuri to tabe ta
leaf TOP koala NOM slowly MAN eat PAST
"The koala slowly ate a **leaf**"

(13) a. コアラが　木　に　い　な　かった
koara ga ki ni i na katta
koala NOM tree DAT to be NEG PAST
"The koala was not in the tree"

b. コアラが　木　に　は　い　な　かった
koara ga ki ni wa i na katta
koala NOM tree DAT TOP to be NEG PAST
"The koala was not in the **tree**"

One way of mapping topicalization into languages without topic marking such as English is via prosodic stress, as indicated by the boldfacing of the topicalized constituent in the translations above.

[26] Written *ha* (は) but pronounced /wa/ but when used as a topic case marker, as noted earlier.

Any constituent can be topicalized, and the topic can also be introduced anew into the clause, i.e. it is possible for the topic to not correspond to any non-topic constituent. For example, while Example (14) is a well-formed Japanese sentence:

(14) スポーツは　サッカーを　やって　い　　る
supōtsu wa sakkā o yat te i ru
sports TOP soccer ACC do TE CONT NONPAST
"As for sports, (I) play soccer"

it is not possible to construct an equivalent sentence with a non-topic marker for *supōtsu* "sports" which means the same as the original (modulo topicalization).

One common function of the topic marker is to contrast certain constituents with other constituents, rather than to mark a true topic (although the distinction between topics and contrastively-marked constituents can be subtle: Kuno (1973); Heycock (2008); Vermuelen (2009)). Returning to our earlier example of zero anaphora in Example (9), e.g., a more natural rendering of the clause would be:

(15) 水　　は　飲ま　な　かった
mizu wa noma na katta
water TOP drink NEG PAST
"(It) didn't drink water"

where *mizu* "water" is marked with the topic marker to contrast it with *happa* "leaf".

Topics tend to occur only in the matrix clause of a sentence (although the topic marker can be used as a contrastive marker in subordinate clauses). There tends to be only one true topic in a sentence, but it is possible for multiple arguments to be marked with the topic marker in contrastive contexts.

In addition to zero anaphora of verbal arguments, it is also possible to elide the predicate in a clause using the copula as a pro-verb, often in conjunction with a topicalized subject. A famous example of this is (in the context of going around a table ordering food at a restaurant, from a customer):

(16) a. 僕　は　鰻　　だ
boku wa unagi da
I TOP eel is
"I (will have) eel" (lit: "I am eel")

In Section 4, we observed that Japanese has a relatively free word order, in that case-marked constituents can be permuted relatively freely. While the core meaning of the clause is unchanged under word order permutation, the information structure of the discourse can change, with the first constituent in the clause receiving focus. Returning to our earlier example from Example (3c), for example (reproduced below as Example (17)):

(17) ゆっくりと　コアラが　葉っぱを　食べた
yukkuri to koara ga happa o tabe ta
slowly MAN koala NOM leaf ACC eat PAST

the focus for this word order is on the fronted adverb and is roughly equivalent to the English "Slowly, the koala ate leaves".[27]

Perhaps one of the best known properties of Japanese is its elaborate system of politeness/formality (Kuno, 1973; Kuno and Kaburaki, 1977). Politeness is an encoding of the relationship between the speaker, the addressee and the referent; formality, on the other hand, is a reflection of the social situation/medium of communication. There are interactions between the two, but also important distinctions. In situations such as discussions between peers with a high degree of familiarity regarding a superior, the formality of the language is often low, but politeness is high when referring specifically to a superior (e.g. a university professor or boss) or their actions. In speech between businessmen from different firms with the intention of forging a long-term relationship, high levels of politeness are used, but formality is often moderate, as over-formality tends to be interpreted as a barrier to intimacy. In technical writing, the language used is highly formal but there is no politeness marking.

In terms of formality, written Japanese has two relatively standardized variants (irrespective of age, gender, etc.): a formal register (e.g. in newspapers or technical publications), and a semi-formal register (e.g. in letters or children's books). Politeness is not marked in formal written Japanese, other than in very rare situations such as when referring to the Japanese imperial household in newspapers. Spoken Japanese covers a much broader spectrum of politeness and formality, and is differentiated based on factors including the age and gender of the speaker, the formality of the situation, and the relationship between the speaker and hearer, and third-party referents.

[27] The effect is roughly equivalent in Example (3a) and Example (3b), but it is hard to recreate the effect of word order variants which reverse the direct object and subject in English; the English cleft construction is often used to convey the impact on information structure, but this tends to over-exaggerate the effect.

Politeness and formality are generally marked based on lexical choice and lexical marking. Pronouns in particular are strong markers of politeness and formality, as well as the gender of the speaker. The singular first person pronoun alone has around a dozen different forms in common use in standard Japanese (and many more in dialects of the language), ranging from *washi* (low politeness, low formality, [older] male speaker; spoken only) and *atashi* (low politeness, low formality, [younger] female speaker; spoken only) to *watashi* (high formality, gender-neutral; spoken and written) and *watakushi* (high politeness, high formality, gender-neutral; spoken and written). In formal written text there is a tendency to avoid using personal pronouns altogether, and use zero anaphors for first person subjects in particular; additionally, expressions such as *hoN* "this" and *tō* "this" are used as a substitute for the adnominal modifier *sono* "this" (e.g. *hoN shuhō* "this method"). Lexical choice of verbs can also be a strong indicator of politeness and formality. The copula verb has a wide range of different forms, encoding different levels of formality — ranging from *da* (low formality) to *desu* (medium formality) to *de ar(u)* (high formality) — and also politeness (see the comments below on empathy and politeness). In general, verbal nouns are more formal than verbs with the same meaning (e.g. *sakusei suru* (作成する) "to make, to create" is more formal than *tsukur(u)* (作る) "make, create").

Lexical marking of politeness and formality takes place primarily on verbs and nouns, and to a lesser extent on adjectives, adjectival nouns and adverbs. In formal written and informal spoken Japanese, verbs are written in **base** or *ru*-form (e.g. *tabe(ru)* "eat") as in all our examples above), whereas in semi-formal and formal spoken Japanese, verbs take the *masu*-form (e.g. *tabe(masu)* "eat"). Nouns vary little with formality (as distinct from pronouns which vary considerably), but can be marked for politeness through marking with prefixes such as *o* or *go* (e.g. *hana* "flower" → *o-hana* "flower", and *kazoku* "family" → *go-kazoku* "family").

The main use of politeness is to codify the relationship between the speaker, hearer and the third-party referee. Perhaps most famously, the choice of the suffix on a name (e.g. the surname *Tanaka*) is a strong marker of politeness/respect towards the referent, from no suffixing (e.g. the bare *Tanaka*) to indicate in-group familiarity[28] to the informal *kuN* for male inferiors (e.g. *Tanaka-kuN*) to the neutral *saN* (e.g. *Tanaka-saN*) and the honorific *sama* for superiors in formal con-

[28]Including referring to an in-group superior with no suffixing in formal contexts where the addressee is an out-of-group individual and high levels of politeness are being used.

texts (e.g. *Tanaka-sama*). For people in high-ranking roles, in spoken Japanese, a name suffix which is indicative of their role/rank is often used (e.g. *seNsei* for teachers, professors, lawyers, politicians, doctors, etc., and *buchō* "department head"). Equally, when using verbs such as "give" and "receive", the speaker is forced to encode their relationship with the givee/receivee, according to: (1) equal status between the speaker and givee (*age(ru)* "(speaker) give (to givee)") or receivee (*mora(u)* "(speaker) receive (from receivee)"); (2) the speaker having lower status than the givee (*sashiage(ru)* "(speaker) give (to givee)"), giver (*kudasar(u)* "(giver) give (to speaker)") or receivee (*itadak(u)* "(speaker) receive (from receivee)"); and (3) the speaker having higher status than the givee (*yar(u)* "(speaker) give (to givee)"). These verbs can be used in literal contexts, but also as auxiliary verbs to mark the metaphoric transfer of a favour through some act, e.g. *tabete age(ru)* "(I do a social equal a favour and) eat" or *tabete itadak(u)* "(I receive a favour from a socially superior person and have them) eat". For a small number of other verbs, there are two lexicalized forms of the **basic** verb (e.g. *ik(u)* "go" or *shabe(ru)* "speak") that express politeness towards the hearer through: (1) the humble form or *keNjōgo*, indicating that the speaker is performing the act (e.g. *mair(u)* "go" or *mōshiage(ru)* "speak"); and (2) the honorific form or *soNkeigo*, indicating that the (socially superior) hearer is performing the act (e.g. *irasshar(u)* "go" or *ossha(ru)* "speak"). For verbs such as *ara(u)* "wash" which have no such lexical variants, the humble form can be formed in combination with *or(u)* "to be" (e.g. *aratte or(u)* "(I) wash", and the honorific form can be formed in combination with *irasshar(u)* "go" (e.g. *aratte irrasshar(u)* "(you) wash"); for verbal nouns such as *shusseki* "attend", the humble form is formed in combination with *itas(u)* "do" (e.g. *shusseki itas(u)* "(I) attend"), and the honorific form is formed in combination with *nasar(u)* "do" (e.g. *(go-)shusseki nasar(u)* "(you) attend"), with the *go-* politeness marker optionally prefixing the verbal noun for extra politeness.[29]

Politeness marking (esp. of nouns) can also be a mark of femininity (e.g. in informal speech, in marking nouns such as *hana* with *o*).

6 Japanese Natural Language Processing

Current research on Japanese natural language processing covers similar topics to other languages, encompassing research on fundamental

[29]It is also possible to prefix the verbal noun with *go-* and use the standard politeness-neutral *su(ru)* light verb. Note that the politeness marking of the verbal noun can only be used for *soNkeigo*, not *keNjōgo*.

issues such as parsing and word sense disambiguation, combined with research on applications such as search and translation.

A few fields stand out as particularly important to Japanese: text-input, segmentation and machine translation. The first two are driven by Japanese's particular orthography, and the last by the fact that there is a huge market for translation, especially between Japanese and English.

6.1 Encoding

Many different characters are necessary to write Japanese: the latin alphabet, hiragana, katakana, and thousands of Chinese characters. Therefore a single byte (with 255 possibilities) is not enough to encode all the characters. Because of this, Japanese, typically uses a multi-byte encoding, where two or more bytes encode a single character. There are several standards for encoding Japanese. The major ones are Shift-JIS, EUC, ISO-2022-JP, UTF-8 and UTF-16. The first three are based on the Japan Industrial Standard (JIS) character sets, the latter two on the Unicode character set.

While Unicode is becoming more common, most Japanese email is encoded using ISO-2022-JP, web pages in Shift-JIS, and mobile phones in Japan usually use some form of Extended Unix Code. Choosing the wrong encoding causes mojibake 文字化け *mojibake* "misconverted garbled/garbage characters, lit: transformed characters" and thus unreadable text on computers. There are excellent discussions of encoding issues online at `http://www.sljfaq.org/afaq/encodings.html` and in Lunde (1999). Here we will merely summarize some of the main differences between the major encodings.

Shift JIS was used by early Microsoft Windows and Macintosh operating systems. It is neither efficient or easy to process. EUC (EUC-JP) is the Unix encoding of JIS. It is relatively efficient for Japanese (most characters can be encoded in two bytes) but does not have a lot of space for non-Japanese — it does not include, for example, latin characters with umlauts (ö) or Korean hangul. ISO-2022-JP-2 is a stateful encoding that allows you to mix different character sets. This means it can represent many different languages, but is slightly complicated to process. It only uses 7 bits of each byte, so is safe even on old 7-bit transfer protocols.

UTF-8 is the Unicode encoding standard widely used in Unix and on the internet, and UTF-16 the Unicode encoding standard in Windows. UTF-8 uses three bytes per kanji, but only one for latin letters, UTF-16 uses two bytes for almost all characters. Depending on the composition of your text, one may be more space efficient than the

other. Unicode covers a much wider range of characters than JIS: most languages can be represented using it. One potential drawback of using a Unicode based encoding is that the same Chinese characters may be represented using different glyphs in Chinese, Japanese and Korean, and the encoding does not say anything about which language is being used.

Most Japanese text processing is done using either EUC-JP or UTF-8 with the latter gradually becoming more common. Recent versions of processing tools such as JUMAN (which used to only work with EUC-JP) now support UTF-8.

In addition to the issues of handling individual characters, Japanese can be written in horizontal style (*yokogaki*: left-to-right then top-to-bottom) like English, or vertical style (*tategaki*: top-to-bottom then right-to-left, standard for novels and newspapers). Punctuation characters are slightly different for the two directions, as follows for the example of 漢字 *kaNji* with traditional quotation marks:

(18)

Horizontal writing	Vertical writing
「漢字」	﹁ 漢 字 ﹂

Typically this is handled by the word-processor, which will have different modes for horizontal and vertical texts, each with different fontsets, the actual characters will be the same vertically or horizontally.

6.2 Text Input

Text input is complicated for Japanese due both to the fact that there are four sets of characters in common use (latin, hiragana, katakana, and kanji) and that there are so many distinct characters. It is infeasible to have a keyboard with all the characters on it, leading to software solutions for text input using a standard-sized keyboard. Development of **front end processor**s (FEP, also known as **input method editors**: IME) led to two main approaches. In one, the Japanese keyboard has kana keys also marked, and one can switch between latin and hiragana/katakana: for example, to type *no* (の) one simply hits the key marked [の] (standardly the [K] key). In the second, kana is entered by its romanized pronunciation: to type の *no*, you would type [N][O] and the FEP would compose them into the single character *no* (の). After this, the FEP can trigger conversion: *no* can be converted to a range of other forms, including katakana, and kanji with the same reading (such as 野, 之, 乃, …). Making Japanese input more efficient was a big

Table 4: Morphological analyser output for 私のナマエは中野です

Word	Pronunciation	Lemma	Part-of-speech
私	watashi	私	noun-pronoun
の	no	の	particle-conjunction
ナマエ	namae	ナマエ	unknown-word
は	wa	は	particle-adverbial
中野	nakano	中野	noun-proper-name
です	desu	だ	copula

research topic in the 1980s, focusing first on creating larger dictionaries, to allow whole words to be entered. Next, was the addition of frequency information, listing entries in order of likelihood. Further advances allowed whole phrases to be entered and disambiguated at once: a famous example was 私の名前は中野です ***w****atashi no* ***n****amae wa* ***n****akano desu*. The Wnn system (named after this example, and developed by Kyoto University and Omrom Corporation) allowed long phrases to be converted in a single pass (Lunde, 1999, Ch. 5). Recent advances now include customization — where a system remembers which words each individual user uses most often — and more complex statistical models based on even larger contexts.

Text input using FEPs is typically interactive: the user types some text, then at a suitable boundary attempts to convert to the correct kana/kanji combination. Text segmentation, on the other hand is normally done fully automatically over a precompiled text, to recover the morphemes as accurately as possible. Segmentation is an essential first step for most natural language processing tasks, including indexing and parsing. Typically systems use large lexicons, augmented with information on parts of speech, frequency and even semantic classes.

6.3 Morpho-syntactic Analysis: Segmentation, Tagging and Parsing

Most Japanese morphological analysers combine the tasks of segmentation, part-of-speech tagging and lemmatization. For example, for the sentence 私のナマエは中野です *watasinonamaewanakanodesu*, we get output such as in Table 4.

One influential and widely used morphological anlayzer is JUMAN (Kurohashi and Nagao, 1998b). This was developed at Kyoto University along with the KNP parser (discussed below). JUMAN uses a large hand-built dictionary, with detailed parts of speech and hand-weighted connections between them. The dictionary is often updated and has been used to segment vast amounts of web text (Murawaki and Kuro-

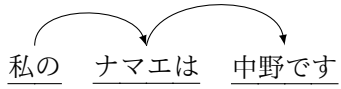

Figure 1: Dependency parse for 私のナマエは中野です; bunsetsu are grouped by underline

hashi, 2010). Instances of NLP applications that use JUMAN to segment text include Chapters 6 and 12. Another popular analyser is ChaSen (Matsumoto et al., 1999), developed at NAIST as part of a suite of NLP software. ChaSen was developed from an early version of JUMAN. Rather than using hand-built connections it uses Hidden Markov Models to learn character and part-of-speech transitions based on word and part-of-speech labels in a training corpus. It can learn models for various dictionaries, including JUMAN's dictionary and IPAL (Information-technology Promotion Agency, 1996). Instances of NLP applications that use ChaSen to segment text include Chapters 2, 3, 9 and 10. ChaSen was forked again into the system MeCab (Kudo et al., 2004), with an improved learning algorithm (conditional random fields) and a faster implementation. JUMAN, ChaSen and MeCab are all open source software, and can be freely downloaded along with their language models.

To recover the syntactic structure between morphemes, we need a parser. The Kurohashi-Nagao-Parser (KNP: Kurohashi, 1998) is a very influential dependency parser. It takes the output of JUMAN, groups the words into bunsetsu, and then links them with dependency relations. Dependency parsers are popular for Japanese as they allow different word order variations to have the same basic structure. For example, the parse for 私のナマエは中野です is shown in Figure 1. JUMAN and KNP were developed in parallel with a large treebank of Japanese: the Kyoto Corpus (Kurohashi and Nagao, 1998a). In this corpus, text from the Mainichi Shinbun corpus (1995) was analysed with JUMAN and parsed with KNP, with the system output being examined and corrected by hand. The corpus has 38,000 sentences and around a million words. Around 5,000 sentences have also been tagged with semantic role labels, zero pronouns and coreference (Kawahara et al., 2002). KNP is used in Chapters 6 and 12.

CaboCha is another popular dependency parser (Kudo and Matsumoto, 2002b) which also chunks words into bunsetsu and then then links them with dependency relations. CaboCha is used in Chapter 13, and was also developed at NAIST using machine learning.

Dependency grammars over bunsetsu do not cover the relationships

between words within the bunsetsu, or allow for grammatical relations such as control where the same word fulfills two roles. More expressive grammars based on Lexical Functional Grammar (Chapter 4) or Head-driven Phrase Structure Grammar (Siegel and Bender, 2002) have also been developed for Japanese.

6.4 Lexicons and Corpora

Morphological analysis and segmentation relies crucially on dictionaries. One of the first widely available lexicons was that from the Information-technology Promotion Agency (IPA). These had detailed syntactic descriptions (Information-technology Promotion Agency, 1996, 1987a) and were used as a base in many systems. Another widely used dictionary was that used by the JUMAN (Kurohashi and Nagao, 1998b) system. They were both similar overall, with some differences as to how they treated some suffixes: in JUMAN, for example, the copula associated with nominal adjectives was treated as an inflection and thus part of the adjective; in the IPADIC it was treated as a separate morpheme. In general JUMAN produced fewer morphemes. Both JUMAN and IPADIC were often inconsistent in their treatment of compositional nouns. For example, 二輪-車 *niriNsha* "two wheeled vehicle" is separated into two morphemes by JUMAN, but left as one by IPADIC. 四輪車 *yoNriNsha* "four wheeled vehicle" is not separated by either.

UniDic (Den et al., 2008) attempts to always consistently split into the smallest possible morpheme. It also adds information about the origin of a word (Native Japanese, Sino-Japanese, other foreign or mixed). Knowing the source helps to improve the accuracy of the segmentation. An example showing the differences in segmentation is shown in Example (19).

(19) 綺麗な四輪車

a. 綺麗 な 四輪車
kirei *na* *yoNriNsha*
pretty COP four-wheeler

b. 綺麗な 四輪車
kireina *yoNriNsha*
pretty four-wheeler

c. 綺麗 な 四 輪 車
kirei *na* *yoN* *riN* *sha*
pretty COP four wheel vehicle

"nice four-wheeler"

Bilingual dictionaries are also useful resources for many tasks, especially translation. Two commonly used ones are EDR and EDICT (EDR, 1996; Breen, 2004). EDR also contains a concept dictionary and corpus, while EDICT contains multiple languages and is open source.

Japanese NLP has also made use of various resources for describing meaning. An early standard is the Bunruigoihyou (Kenkyujo, 1964): a flat five-level classification of meanings covering some 55,000 nouns (see Section 2.1 on page 112 for a fuller description). It is used in Chapters 6 and 7. Other popular resources are GoiTaikei: a Japanese Lexicon (Ikehara et al., 1997, used in Chapter 12), which also has verb semantic preferences; and the Japanese Wordnet (Isahara et al., 2008) which links meanings to wordnets in many languages and has an accompanying sense-tagged corpus (Bond et al., 2012).

In addition to the EDR corpus, there are several corpora in wide use. Perhaps the earliest was the ATR corpus , which had transcribed dialogs of travel conversations (reserving hotel rooms) in both English and Japanese, with segmentation and part of speech tags (Morimoto et al., 1994). This was later extended with a much larger collection of travel expressions from phrase books: the BTEC corpus (Takezawa et al., 2002).

Another influential corpus was the Kyoto Corpus (Kurohashi and Nagao, 1998a). This consists of 38,000 sentences and roughly a million words. The first half comprises seventeen days of the Mainichi Shinbun, from 1995. The remainder was all the editorials from that year. It was originally tagged with JUMAN and parsed with KNP. It was then retagged with the IPA tags and used to train ChaSen. Other projects have tagged it with different data, such as predicate argument structure for verbs, adjectives and event nouns, and coreference information (the NAIST Text Corpus: Iida et al., 2007). It has also been tagged with senses from Lexeed and GoiTaikei as part of the Hinoki Corpus (Bond et al., 2008) as well as translated into Chinese and English (Uchimoto et al., 2004).

The National Institute for Japanese Language and Linguistics (NINJAL) is producing a series of corpora in the KOTONOHA project.[30] These include the Balanced Corpus of Contemporary Written Japanese (BCCWJ), the Taiyō Corpus, and the Corpus of Spontaneous Japanese (CSJ). In addition, they are currently compiling a corpus of historical Japanese and a very large corpus of modern Japanese (one trillion words).

BCCWJ is a balanced corpus of one hundred million words of con-

[30] http://www.ninjal.ac.jp/english/products/kotonoha/

temporary written Japanese. There are three subcorpora: a random selection of all books, magazines, and major newspapers published in the years 2001-2005; all books that are catalogued at more than 13 metropolitan libraries in Tokyo; and a collection of mini corpora selected for specific research purposes of the NINJAL research groups (such as governmental white papers, textbooks, laws, bestselling books, and web text). The corpus is automatically segmented and POS tagged with two layers: short unit words (similar to Unidic) and long unit words (similar to IPADIC).

The Taiyō Corpus consists of texts from the periodical Taiyō. There are 3,409 articles in 60 issues published over the period of 1895-1925, with a total of 15 million characters. The articles show many different writing styles and orthographic variations.

The Corpus of Spontaneous Japanese (CSJ) consists of high quality recordings of 650 hours of spontaneous speech (about 7 million words). There are 1,400 different speakers with ages from 20-90. 95% of the CSJ is devoted to spontaneous monologues, such as academic presentations and public speaking. 5% consists of spontaneous dialogues and reading aloud. The corpus is well annotated, with transcriptions, parts of speech, labels of phonetic segmentation and intonation.

Finally, as researchers in Japan realize the importance of making resources **accessible** as well as **useful** (Ishida, 2006) there have been several open source corpora released. Many of these are multilingual, including the Tanaka Corpus (Tanaka, 2001) with around 150,000 sentence pairs; the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles[31] with around 500,000 sentence pairs, and the English-Japanese Translation Alignment Data[32] which is partly described in Chapter 9.

6.5 Machine Translation

Machine translation research has always been a big topic in Japanese NLP. Because Japanese and English are so different linguistically, translation is difficult, with zero pronouns, different word orders and significant differences in what is marked in the two languages (for example, Japanese marks politeness, while English marks number and definiteness). Early research concentrated on syntactic or semantic transfer: the source language was parsed to some more abstract representation (such as a dependency parse, phrase structure tree or case frame), this was transferred to the target language, and then the target string generated, as in Figure 2. Some systems use dependencies as the

[31] `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`
[32] `http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/`

Figure 2: Transfer based Japanese-English machine translation

representation (Nakazawa and Kurohashi, 2008), some use syntactic trees, like the system outlined in Chapter 11, some use case-frames, such as **ALT-J/E** (Ikehara et al., 1991) and some use deeper representations such as Minimal Recursion Semantics (Bond et al., 2011).

Because of the vast number of translation divergences between English and Japanese, Japanese NLP researchers pioneered work to learn translations from examples, in the form of example-based machine translation (EBMT: Nagao, 1984). More recently, research on machine translation involving Japanese has moved to include statistical machine translation (Brown et al., 1993; Yasuda et al., 2010).

A recent addition to Japanese NLP resources is the Natural Language Tool Kit (NLTK), an introduction to NLP using the Python language that comes with extensive open-source code (Bird et al., 2009). There is a complete Japanese translation of the NLTK book that has a full extra chapter on Japanese NLP (Bird et al., 2010). The English book, with a translation of the Japanese chapter, is available on-line: `http://nltk.org/book/`.

In addition to the resources described here, up-to-date lists of resources related to Japanese NLP can be found at the following sites:

- The web page of the **Association for Natural Language Processing** in Japan has information about the society's meetings and Journal along with a list of links to related information.
  `http://www.anlp.jp/`
- The **Association for Computational Linguistics** (ACL) has a list of resources for Japanese (and many other languages), including corpora and tools.
  `http://aclweb.org/aclwiki/index.php?title=Resources_for_Japanese`
- **Natural Language Processing Portal Site** is produced by the Knowledge Information Processing Technologies Committee of JEITA (Japan Electronics and Information Technology Industries Association). It has perhaps the most comprehensive list of Japanese resources and tools (mainly in Japanese), and includes links to pa-

pers using the resources.
http://www.jaist.ac.jp/project/NLP_Portal/doc/LR/lr-cat-e.html

- **Advanced LAnGuage INformation Forum** (ALAGIN) brings together representatives of industry, academia and the government to research, develop, test, and standardize text and speech translation systems, spoken dialogue systems, information retrieval and analysis technology. The forum also develops and distributes linguistic resources (dictionaries, corpora, etc.) for use in these systems.
http://www.alagin.jp/index-e.html
- **Gengo Shigen Kyōkai** (GSK) "Language Resource Association" is a non-profit organization for promoting the distribution of language resources such as speech data, lexicons, text corpora, terminology, and various tools for language processing, primarily for Japanese.
http://www.gsk.or.jp/en/

7 Overview of the book

This book is aimed at people interested in natural language processing involving the Japanese language. It introduces twelve papers on some of the classic problems in Japanese NLP:

The first part of this volume deals with morphology and syntactic analysis. Chapters 2 (*Domain-Specific Statistical Data for Morphological Analysis*) and 3 (*Detecting Japanese Term Variation by Morpho-syntactic Rules*) deal with morphological analysis, especially the problem of unknown words. Chapter 4 (*Construction of a Japanese Parsing System based on LFG*) presents a sketch of an implemented linguistically-precise grammar of Japanese using LFG.

The second part of this volume looks at issues relating to discourse in Japanese. Chapter 5 (*Dialogue Translation Method using Participants' Social Roles*) shows how properly analysing the lexicalization of politeness in Japanese improves the quality of machine translation. Chapters 6 (*Statistical Anaphora Resolution for Japanese Zero Pronouns*) and 7 (*Translation of Pronouns in E-to-J Machine Translation*) deal with problems of identifying and generating zero pronouns. Finally, Chapter 8 (*Processing Japanese Self-repair in Spoken Dialogue Systems*) deals with problems of repetition and repair in spoken dialogue.

In the third part of the volume, we present applications relating to Japanese NLP. The first three chapters deal with translation. In Chapter 9 (*Measures for Aligning J-E News Articles and Sentences*), trans-

lation data is automatically aligned so that it can be used as training data for machine translation. Chapter 10 (*Balancing up Efficiency and Accuracy in Translation Retrieval*) looks at translation memories and lexical similarity, with the surprising result that character-based indexing consistently outperforms word-based indexing. Chapter 11 (*Hierarchical Phrase Alignment Harmonized with Parsing*) shows a method of learning transfer rules from aligned text. Chapter 12 (*Paraphrasing Predicates based on Case Frame Alignment*) aligns predicates in dictionaries with their definitions to produce rules for monolingual paraphrasing. Finally, Chapter 13 (*Sentence Reconstruction in Summary Generation*) investigates how text is rewritten when it is summarized, aligning dependency parses and investigating the differences.

We hope that you find these papers interesting and informative.