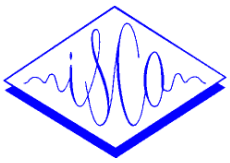


ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

Japanese copula marker works as a filler in spontaneous speech

Takehiko Maruyama † ‡ Hideki Tanaka † ‡ Hideki Kashioka †

† ATR Spoken Language Translation Research Laboratories

‡ The National Institute for Japanese Language

Abstract

The Japanese syntactic form *desune* generally works as a copula marker, but sometimes it also works as a filler in spoken language. We examine the distribution and characteristics of two aspects of *desune* through spontaneous speech corpora, called ASU and CSJ. We also compare *desune* with other fillers and indicate the similarities between them.

1 Introduction

Since various kinds of spontaneous speech corpora have become available in recent years, we are now able to process spoken language for a variety of purposes. It is now common knowledge that spoken language has many distinctive features in comparison with written language, however, there have been very few linguistic studies conducted on the differences between the two. In this paper, we will discuss the characteristics of the Japanese copula marker *desune* in spoken language, especially its peculiar usage as a filler.

Desune is a syntactic constituent[1] that essentially works as a copula marker (CM) at the end of a sentence, as in (1).

(1) *kore-wa Ken-ga kaita hon desune.*
this-TOP Ken-NOM write-PAST book CM
'This is the book Ken wrote.'

However, especially in spontaneous speech, *desune* can be placed inside a sentence. We can regard the *desune* in (2) not as a copula marker, but as a filler (FIL) — a freely adjoined element inside a sentence.

(2) *Ken-wa desune hon-o desune kakimashita.*
Ken-TOP FIL book-ACC FIL write-PAST
'Ken wrote a book.'

These two aspects of *desune* cause problems for sentence boundary detection in ASR, parsing and translating a sentence in NLP. In each case, we need to detect whether an instance of *desune* works as a copula marker or as a filler.

[1] Morphologically, *desune* is composed of copula marker '*desu*' and sentence-final particle '*ne*.' We describe it *desune* for short.

We classify *desune* into two groups by its position: **ESD** (End of Sentence *Desune*) and **ISD** (Inside a Sentence *Desune*). The former occurs at the end of sentence, preceding a period, and works as a copula marker, as in (1), whereas the latter occurs inside a sentence and works as a filler, as in (2).

In this paper, we examine the distinctive features between two usages of *desune*, and investigate their distribution through the spontaneous speech corpora.

2 Corpus I: ASU

Our first target is the spontaneous monologue corpus called "*Asu-wo-Yomu* (**ASU**)." ASU is the transcription of a ten-minute TV commentary program from the Japan Broadcasting Corporation (NHK), in which a commentator talks about political, economic, and social issues. Although the manuscript for the lecture has been carefully prepared, spontaneous speech can often be observed.

We recorded 300 programs of ASU on video, transcribed all the utterances, including fillers, into text, and punctuated manually. We also tagged part-of-speech (POS) labels of every morpheme[2]. The summary of the corpus is shown in Table 1.

Table 1: Summary of ASU

programs	300
speakers	40
sentences	18,163
morphemes	532,107
mor/sentence	29.3

3 Analysis

3.1 Frequency

First, we show the frequency of *desune* in ASU. We counted the total occurrences of *desune*, and classified them into ISD and ESD. Table 2 shows the distribution of the two. The ratio of ISD accounts for more than half (63%), meaning that *desune* tends to occur inside a sentence more frequently than at the end of a sentence.

[2] POS labels were tagged manually in the style of TDMT, Transfer-Driven Machine Translation system. See [1].

Table 2: Distribution of ISD and ESD (ASU)

Total	ISD	ESD
412 (100%)	259 (62.9%)	153 (37.1%)

3.2 Difference between speakers

We counted the frequency of ISD and ESD in each program, and correlated the results with the speakers to investigate whether there is a speaker-oriented difference in the use of *desune*. Table 3 shows the top fifteen rankings.

Table 3: Frequency of *desune* in one program.

Total	ISD	ESD	Speaker (charge)
48	34	14	Speaker A (5)
41	33	8	Speaker A (5)
39	36	3	Speaker A (5)
38	34	4	Speaker A (5)
33	26	7	Speaker A (5)
21	12	9	Speaker B (1)
13	8	5	Speaker C (9)
11	3	8	Speaker D (3)
9	5	4	Speaker E (1)
7	4	3	Speaker F (1)
7	3	4	Speaker G (1)
7	0	7	Speaker C (9)
6	4	2	Speaker C (9)
6	3	3	Speaker H (4)
6	1	5	Speaker D (3)

The greatest frequency was 48 utterances by speaker A who appeared on ASU five times. A's programs were ranked in the top fifth, which means he uses *desune* with very high frequency, especially inside sentences.

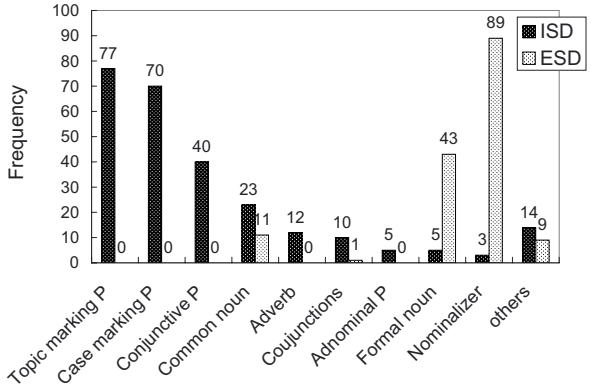
The total occurrences of *desune* appearing in the fifteen highest ranks amounted to 292, which covered 71% of total frequency of *desune* (412). On the other hand, there were sixteen speakers who didn't use *desune* at all, and 221 programs in which no *desune* occurred. From the practical viewpoint, we can infer that *desune* is used by a small number of speakers and its occurrence depends on speakers' individual speaking styles.

3.3 Occurence positions

Here, we examine the occurrence positions of *desune*, investigating the morphemes preceding ISD and ESD. At first, we show the distribution of POS preceding ISD and ESD in Figure 1. The 'P' stands for 'Particle'.

The distribution of POS preceding ISD and ESD is completely different. In ISD, three types of particles —topic markers, case markers, and conjunctions— are more prominent than other elements. In ESD, on the other hand, the nominalizer shows conspicuous frequency, and formal nouns follow. Common nouns can be observed in both ISD and ESD.

Figure 1: POS preceding ISD and ESD (ASU)



Tables 4 and 5 show the top ten rankings of the morphemes preceding ISD and ESD.

Table 4: Morphemes preceding ISD (ASU)

	Frequency		Form	(POS)
	68	(26.3%)	*wa*	(Topic marking P)
	26	(10.0%)	*te*	(Conjunctive P)
	19	(7.3%)	*ni*	(Case marking P)
	14	(5.4%)	*de*	(Case marking P)
	14	(5.4%)	*o*	(Case marking P)
	13	(5.0%)	*ga*	(Case marking P)
	9	(3.5%)	*mo*	(Topic marking P)
	9	(3.5%)	*to*	(Conjunctive P)
	5	(1.9%)	*kara*	(Case marking P)
	4	(1.5%)	*toka*	(Adverbial P)
Total	259	(100%)	72 kinds	

Table 5: Morphemes preceding ESD (ASU)

	Frequency		Form	(POS)
	89	(58.2%)	*n*	(Nominalizer)
	21	(13.7%)	*wake*	(Formal noun)
	18	(11.8%)	*koto*	(Formal noun)
	3	(2.0%)	*mono*	(Formal noun)
	2	(1.3%)	*toka*	(Adverbial P)
	1	(0.7%)	*teikyou*	(Verbal noun)
	1	(0.7%)	*shidai*	(Formal noun)
	1	(0.7%)	*mijikai*	(Adjective)
	1	(0.7%)	*amari*	(Adjectival noun)
	1	(0.7%)	*teki*	(Suffix)
Total	153	(100%)	26 kinds	

In ISD, various particles occurred before *desune*, especially the topic marker *wa* (26.3%). The total frequency of the top ten (181) accounts for 75% of all (259) ISD. In ESD, on the other hand, the nominalizer *n* amounts to more than half (58.2%), and no particle occurred before *desune*. The total frequency of the top three (127) accounts for 84% of all (153) ESD. We can conclude that there is a clear distinction between ISD and ESD from the point of view of occurrence positions.

4 Discussion

Based on the observations so far made, we will discuss the characteristics of morphemes preceding *desune*. The subjects here are particles and common nouns.

4.1 Particles preceding desune

As we have observed, particles like *wa*, *ni*, *de*, *o*, *ga*, and *kara* occurred before ISD only, and never before ESD. We can partly prove this result from the grammatical account. If the particles *wa*, *o*, and *ga* are placed before ESD, the sentences are completely ungrammatical.

(3) **keeki-o* *yaita-no-wa* *Naomi-wa* *desune.*
cake-ACC baked-TOP Naomi-TOP CM
'It is Naomi who baked the cake.'

(4) **kono kabin-o* *watta-no-wa* *Ken-ga* *desune.*
this vase-ACC broke-TOP Ken-NOM CM
'It is Ken who broke this vase.'

The appropriateness of our observation that *wa*, *o*, and *ga* never occurred before ESD can be explained by this grammatical constraint.

Meanwhile, when the particles *ni*, *de*, and *kara* are placed before ESD, the sentences become grammatical.

(5) *Ken-ga* *kabin-o* *watta-no-wa* *koko-de* *desune.*
Ken-NOM vase-ACC broke-TOP here CM
'It is here where Ken broke the vase.'

(6) *geemu-o* *hajimeru-no-wa* *kore-kara* *desune.*
game-ACC start-TOP now-START CM
'It is from now that we start the game.'

We could not find any examples like these at all in ASU. This result, however, does not mean that the grammatical prediction is not appropriate; rather, it is due to the problem of corpus sparseness and the particular speaking style of ASU for a TV commentary program. Examples like (5) and (6) may be found as we investigate other corpora (See Chapter 5).

4.2 Common nouns preceding desune

We have observed that common nouns occurred before both ISD and ESD in Figure 1. The observed common nouns preceding ISD were mostly modifiers that function as adverbs, as in (7).

(7) *naze* *ima* *desune* *nihon-no* *dentou-ongaku-no*
why now FIL Japanese tradditional music-GEM
taisetsusa-ga *towarete* *iru* *no* *deshou* *ka.*
importance come into question
'Why does Japanese traditional music come into question now?'

The common noun *ima* modifies VP as an adverb, and *desune* follows the modifier as ISD.

On the other hand, there were various kinds of common nouns preceding ESD, like "atmosphere, power, Democratic Party of Japan, care manager," and so on.

(8) *koko-wa* *chikettouriba-desu-ga* *marude*
here-TOP ticket center, but just like
kuukou-no-youna *fun'iki* *desune.*
like an airport atmosphere CM
'Here is the ticket center, but its atmosphere is like an airport'

From the viewpoint of syntax, any common noun can occurr before ESD, whereas common nouns preceding ISD are restricted to those that can function as a modifier alone, like *ima* in (7). Arguments without case-marking particles can also precede ISD, as in (9).

(9) *watashi* *desune* *kono hon-o* *kaita n desu.*
I-ϕ FIL this book-ACC wrote
'I wrote this book.'

The examples discussed above are all within the scope of a grammatical account. However, we found some examples beyond the grammatical estimation, as in (10).

(10) *sougaku-to-wa* *koyou* *hataraiteiru hito-no*
total-TOP employment working people-GEM
kazu *desune* *sorekara* *chingin* *soshite* *jikan...*
number and wages and hours
'Total (means) the employment, I mean the number of workers, wages, and working hours...'

The structure of (10) is complicated because a sentence '*hataraiteiru hito-no kazu desune*' is inserted into another sentence '*sougaku-to-wa koyou sorekara chingin soshite jikan....*' The inserted sentence completed by *desune* is a supplementary explanation for *koyou* in the main sentence. How should we treat this *desune* in the inserted sentence — as ISD or ESD? To solve this problem, we should precisely define the sentence boundaries in spontaneous speech[3], but we will not explore this issue further in this paper. Following our definition of ISD and ESD, we just count *desune* in (10) as ISD, since no period follows.

4.3 Comparing with fillers

As we noticed, ISD, an adjoined element, works as a filler. To compare ISD and other fillers, we examined the occurrence positions of fillers in ASU. Table 6 is the top ten rankings of morphemes preceding all fillers.

Noteworthy here is that the order of morphemes in the rank is almost the same as in the case of ISD, shown in Table 4. *Wa*, *te*, *ni*, *ga*, *o,* and *de* are shared between the two, which means that ISD and other fillers have great similarity in the distribution of the occurrence position. In this perspective, ISD can be regarded as having the same characteristics with other fillers.

Table 6: Morphemes preceding fillers (ASU)

Frequency		Form	(POS)
2,073	(11.6%)	*wa*	(Topic marking P)
1,691	(9.4%)	*no*	(Adnominal P)
1,044	(5.8%)	*te*	(Conjunctive P)
977	(5.4%)	*ni*	(Case marking P)
957	(5.3%)	*ga*	(Case marking P)
671	(3.7%)	*o*	(Case marking P)
558	(3.1%)	*de*	(Case marking P)
438	(2.4%)	*to*	(Case marking P)
408	(2.3%)	*mo*	(Topic marking P)
396	(2.2%)	*ga*	(Conjunctive P)
Total 17,930	(100%)	1,686 kinds	

5 Corpus II: CSJ

In this section, we examine the distribution of ISD and ESD through our second target called CSJ, The Corpus of Spontaneous Japanese[2]. The analysis of distinctive features between ISD and ESD in CSJ is important for the automatic utterance segmentation[3] because CSJ contains no punctuation.

We extracted 2,527 *desune* utterances from 29 monologues[3], and manually determined ISD or ESD. Table 7 shows the distribution of the two. The ratio of ISD accounts for almost 75%, larger than the case of ASU shown in Table 2.

Table 7: Distribution of ISD and ESD (CSJ)

Total	ISD	ESD
2,527 (100%)	1,835 (72.6%)	692 (27.4%)

Tables 8 and 9 show the top ten rankings of the morphemes preceding ISD and ESD. We can see that the distribution of morphemes preceding each *desune* is very similar to the case of ASU, shown in Tables 4 and 5. There are many shared common morphemes: *wa*, *te*, *ni*, *ga*, *o,* and *de* in ISD, and *n*, *wake*, *koto* and *mono* in ESD. We can conclude that the distribution of morphemes preceding ISD and ESD is generalized, even among different types of corpora.

The result of two different corpora showing almost the same tendency for the behavior of *desune* proves that our analysis of ASU can be satisfactorily accepted as a general observation[4]. Although we need to conduct more detailed research and verification in each case, these descriptions can be generalized as linguistic research of spoken language, and will be useful for the practical processing, for example, the automatic segmentation of spontaneous utterances.

[3] We selected monologues including more than 40 occurrences of *desune*.

[4] In addition, the exapmles in which particles *ni*, *de*, *kara* occurring before ESD were observed in CSJ, which we could not find in ASU.

Table 8: Morphemes preceding ISD (CSJ)

Frequency		Form	(POS)
302	(16.5%)	*wa*	(Topic marking P)
270	(14.7%)	*te*	(Conjunctive P)
133	(7.2%)	*ni*	(Case marking P)
124	(6.8%)	*ga*	(Case marking P)
105	(5.7%)	*de*	(Case marking P)
91	(5.0%)	*toka*	(Adverbial P)
83	(4.5%)	*mo*	(Topic marking P)
81	(4.4%)	*o*	(Case marking P)
76	(4.1%)	*to*	(Conjunctive P)
74	(4.0%)	*no*	(Adnominal P)
Total 1,835	(100%)	144 kinds	

Table 9: Morphemes preceding ESD (CSJ)

Frequency		Form	(POS)
201	(29.0%)	*n*	(Nominalizer)
97	(14.0%)	*wake*	(Formal noun)
63	(9.1%)	*koto*	(Formal noun)
10	(1.4%)	*mono*	(Formal noun)
9	(1.3%)	*tokoro*	(Formal noun)
8	(1.2%)	*kata*	(Suffix)
7	(1.0%)	*baai*	(Common noun)
4	(0.6%)	*nai*	(Auxiliary)
4	(0.6%)	*ooi*	(Adjective)
4	(0.6%)	*rashii*	(Auxiliary)
Total 692	(100%)	239 kinds	

6 Conclusion

We examined the characteristic of Japanese copula marker *desune* placed inside a sentence and at the end of a sentence through two spontaneous speech corpora, ASU and CSJ. Similar distribution of the occurrence positions were observed, and distinctive grammatical feartures of preceding morphemes were inferred. We also indicated the similarities of occurrence positions between *desune* placed inside a sentence and fillers.

References

[1] O. Furuse, K. Yamamoto, S. Yamada. "Using Constituent Boundary Parsing for Multi-lingual Spoken-language Translation." *Journal of Natural Language Processing*, **6**(5), 63–91, 1999.

[2] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. "Spontaneous Speech Corpus of Japanese." *Proc. of LREC 2000 2nd International Conference on Language Resources and Evaluation,* Athens, 947–952, 2000.

[3] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. "Identification of "Sentence" in Spontaneous Japanese." This volume.

Note: This research was supported in part by the Telecommunications Advancement Organization of Japan.