AN ANALYSIS OF THE JOINT VENTURE JAPANESE TEXT PROTOTYPE AND ITS EFFECT ON SYSTEM PERFORMANCE

Steve Maiorano
Office of Research & Development
Washington, D.C. 20505
email: maiorano@crl.nmsu.edu

BACKGROUND

The TIPSTER Data Extraction and Fifth Message Understanding Conference (MUC-5) tasks focused on the process of data extraction. This is a procedure in which pre-specified types of information are identified within free text, extracted, and inserted automatically within a template. Three TIPSTER contractors -- BBN, GE/CMU, NMSU/Brandeis -- participated in the August '93 MUC-5 evaluation for both the English joint venture (EJV) and English microelectronics (EME) domains and their Japanese-language counterparts, the JJV and JME applications. Two other contractors -- SRI and SRA -- participated in the EJV and JJV domains alone. CMU's Textract system took part in the Japanese-language domains only. Of the five systems that tested in both English and Japanese, all but one scored higher in the Japanese-language applications according to both the summary error-based scores and recall/precision-based metrics. This overall result has lead some participants and observers to suggest that Japanese is an "easier" language than English.

Japanese-language usage in the total 1297-article JJV corpus exhibits the same degree of ellipsis-generated vagueness and ambiguity as in other domains and genres of Japanese writing. On the other hand, however, in matters of information presentation JJV articles are very formulistic. This paper argues that the stereotypical structure of the *topic sentence* in the JJV corpus together with the "default" pattern of certain template fills gives the Japanese systems a ready basis for extracting information and inserting it into a template. The result is better overall systems' performance in JJV than EJV as indicated by the scoring metrics.

METHODOLOGY

The argument outlined in this paper is based upon a *discourse analysis* of two portions of the entire 1297-article JJV corpus: the 150-article JJV test set and 100 randomly selected development-set articles. In addition, a *descriptive analysis* was performed on approximately 50 JJV test articles and corresponding template results for varying combinations of the six systems that participated in MUC-5; all six systems, however, were analyzed on a subset of 12 selected articles, or a total of 72 individual template results. The entire descriptive examination is motivated by a desire to understand better the various systems' capabilities in order to make the numerical results more tangible to potential users. The assumption is that one can construct a composite performance-based description for each system derived from the analysis of individual templates, and that the resulting snapshot -- what the system actually does -- will be more comprehensible to users than the theoretical model of a system outlined in a technical summary -- what it should do.

Although the *discourse analysis* has not yielded a full-blown discourse structure for the JJV corpus, the most essential element of the evolving top-down paradigm, the topic sentence, is identified. Any attempt to formulate a complete discourse paradigm for JJV must first deal with this sentence. It contains much information significant in its own right and -- more to the point for data extraction -- relevant to template insertion. In fact, **most** of the time the topic sentence contains all the minimally required data for instantiating and tracking a tie-up relationship.

This paper first examines the stereotypical nature of this topic sentence -- hereafter referred to as an article's "Impact Line" -- before moving onto a discussion of the "default" mechanism. The Impact Line prototype operating in conjunction with the instantiation of certain high-percentage slot fills ("defaults") provides a proficient extraction heuristic and corresponding salubrious quantitative effect upon system performance.

JJV DOMAIN AND THE IMPACT LINE

The JV application focuses on tracking tie-ups between at least two entities. It is necessary, therefore, to 1) identify the entities engaged in some business activity or development project and 2) to confirm that the arrangement between them is a tie-up relationship. Therefore, for the Impact Line to have any "impact" at all in this application, its prototype should at least contain the information necessary in fulfilling the above criteria.

Two definitions of the prototypical Impact Line, version 1 and version 2, are presented below. Version 1 discusses the data items necessary to meet the above-mentioned criteria for generating a tie-up: two entities and the indication of a tie-up. In order to show how the structure of this version-1 Impact Line facilitates the identification and extraction of these data items, moreover, the first definition discusses the grammatical role of the Japanese topic marker は "wa," its importance in marking relevant proper nouns in the JJV corpus, and the Impact Line's verbal element. By this definition, 81% of the JJV test set is Impact Line prototypical.

Version 2 is a more restrictive definition requiring the presence of two more extractable data elements in the Impact Line in addition to the criteria of version 1. The second definition, therefore, discusses the types and distribution of Impact Line data items. This version of the prototype occurs 65% of the time.

DEFINITION OF THE PROTOTYPICAL IMPACT LINE (VERSION 1)

(1) IMPACT LINE TOPIC MARKER (GRAMMATICAL FORCE)

In the same way that the Impact Line is crucial to developing a complete discourse paradigm for JJV, or perhaps any domain of Japanese newspaper articles,[1] any discussion about what constitutes a prototypical Impact Line must start with the Japanese topic marker (<TM) "wa" whose role as designator of the Impact Line's grammatical

[1] I am just beginning to analyze newspaper "announcement" articles in other domains, such as JME, to see if the Impact Line prototype has validity and can form the basis for a metamodel that is not domain specific.

subject is predominant in the JJV test corpus. The "wa"-designated subject sets the tone for the Impact Line as the Impact Line does for the JJV article.

In Japanese discourse generally, "wa" is a particle that indicates the theme or topic of a sentence and as such often, but not always, corresponds to the subject of the sentence. Perhaps just as often "wa" serves to highlight or topicalize other pieces of information, while the particle "ga" marks the subject. For example:

Kono hon *wa* Ken **ga** yonda.
(Speaking of this book, Ken has read it.)

Eigo *wa* Ken **ga** umai desu.
(With regards to English, Ken is skillful.)

The subject Ken is designated by **ga** and the topic by *wa.* However, when the subject or agent of the action is also the sentence topic, **wa** marks the grammatical subject. For example:

Ken **wa** kono hon o yonda.
(Speaking of Ken, he read this book.)

It is this latter grammatical function of "wa" as the sentence topic **and** agent-of-action designator that predominates in the JJV test articles. **Example 1** below is #2630 from the JJV test set:

東京海上火災	は	十七日
PN-Subject	<TM	Numeral+N
Tokyo Marine & Fire		17th

イギリス	の	大手総合保険
N	Prt	NP
English		big/gen'l/insur

会社	コマーシャル・ユニオン
N	PN
comp.	Commercial Union

社	(本社・	ロンドン)	と
N	N	PN	Prt
comp.	hqs.	London	with

業務提携した	と	発表した。
VP	Prt	VP
business/tie-up/did		announcem't/did

Translation:
Tokyo Marine & Fire [Insurance Co.] announced on the 17th that it has concluded a business tie-up with a large English general insurance company, Commercial Union (headquarters London).

Given the grammatical importance of "wa" in indicating the subject of the Impact Line, this function takes on added significance in the JV domain where the identification of tie-up entities in a tie-up relationship triggers the extraction process. The Impact Line topic marker in JJV articles is a reliable designator of proper nouns that are valid tie-up partners to be extracted and inserted into the template. In fact, in 117 Impact Lines out of 145[2] JJV test-set articles (81%), "wa" marks at least one tie-up partner;[3] and this tie-up partner is not simply the Impact Line topic, but the agent of action as well.

Furthermore, in 19 instances out of those 117, the topic marker is

[2] Five of the 150 test-set articles produced a template but not any tie-ups because they were about either sister-city relationships or talks that were broken off. Therefore, the baseline figure that will be used hereafter in discussing the JJV test set is 145.

[3] There was a similar high percentage of 79% for 100 randomly selected JJV development set articles.

preceded immediately by two proper nouns designating two principal tie-up partners. Typically the structure will look like **Example** 2 below:

```
(Ex.2) 日本アイ・ビー・エム  と
            PN           Conj
       Japan/IBM         and

住友電気工業          は
   PN              <TM
Sumitomo Electric
```

The conjunction と ("to") binds the two entities IBM Japan and Sumitomo Electric as co-subjects. Alternately this paradigm allows for modifiers before either or both of the entities (**Examples** 3 -- 5):

```
(Ex.3) トヨタ と 米国自動車
       Toyota and US car-

メーカーGM     は
maker GM     <TM
```

```
(Ex.4) 日本の自動車メーカー
        Japanese carmaker

トヨタ  と  GM   は
Toyota and GM  <TM
```

```
(Ex.5)
日本の自動車メーカートヨタ    と
Japanese carmaker Toyota and

米国自動車メーカーGM        は
US carmaker GM          <TM
```

Thus far, the prototypical Impact Line can be encapsulated in the following short notation:

.....X **wa**

where X is a principal tie-up entity and the ellipsis marks allow inclusion of multiple subjects as shown in **Examples** 2 -- 5. It is important to note, moreover, that whether modifiers precede an ENTITY-designate or not, or whether a conjunction is present or not, the topic marker "wa" is preceded immediately -- in the grammatical sense -- by an entity that is a principal tie-up partner. Twenty-one of the 117 "wa"-designated entities are preceded immediately by information about the entity -- such as location -- enclosed in parentheses, rather than the entity name itself. For example:

```
日興証券  (本社・東京)      は
Nikko Securities (hqs. Tokyo) <TM
```

Orthographically this may be misleading, but grammatically the topic marker indicates the entity, not its headquarters location. Therefore, such cases retain their prototypical validity.

(2) IMPACT LINE TOPIC MARKER (PRACTICAL FORCE)

The Impact Line topic marker exerts a force that extends beyond the scope of a JJV article's first sentence. In instances of ellipsis, which occurs frequently throughout the JJV corpus, the appropriate subject can be supplied by inserting the Impact Line "wa"-designated subject. Article #1747 is a classic example of Japanese presentation:

１）　常陽銀行**は**六日、　野村証
券と包括的な業務提携を結ん
だと発表した。２）証券分野ではすで
に日興証券と提携しているが、
多様な地域顧客のニーズに対応
するため複数の会社と提携して
証券サービスを充実させる．．．．
４）野村証券との提携ではM&A（企
業の合併・買収）業務を盛り
込んでいる、常陽銀行**は**この分野
にも積極的に取り組む構えだ。

Literal translation ([ ] indicates zero anaphora):

1) On the 6th Joyo Bank announced

that [ ] had concluded a comprehensive business tie-up with Nomura Securities. 2) In the securities area, [ ] already has a tie-up arrangement with Nikko Securities, but in order to meet the diverse needs of [ ] regional customers, [ ] is making up for the lack of securities-related services through tie-ups with several companies.... 4) As far as the tie-up with Nomura is concerned, M & A (company mergers and acquisitions) business is included, and Joyo is poised to move aggressively into this area.

Note that the Impact Line subject, Joyo Bank, does not appear again until the fourth sentence, which is the last line of the article. Until it reappears as the subject, it is omitted and one needs to supply a pronoun or proper name -- "it", "its", "Joyo" -- in order to read the passage understandably in English. In other words, the heuristic, which states that ellipsis can be filled by the subject marked by the Impact Line topic marker, works quite well here.

Admittedly this is an easy case because stylistically Japanese allows ellipsis in a sentence that follows one in which the subject was introduced originally. In fact, using the term heuristic *qua* a convention with grammatical and stylistic acceptability may be inappropriate. However, in numerous other instances when convenience dominates and ellipsis is propagated throughout a text beyond the decent bounds of style, assigning the proper subject is less clear-cut. Particularly troublesome are those cases in which ellipsis continues for several sentences before the introduction of a new subject appropriately designated by another topic marker. Thereafter, the subject -- which one? -- is again omitted, and one must decide between calling upon the proximate "wa"-designated subject or the original Impact Line "wa"-designated agent.

When coding or checking 100 of the 150 test-set articles, I noted only one instance (#2111) in which context demanded that the subject of a particularly complex sentence was not the default Impact Line "wa"-designated one. It is, therefore, a powerful heuristic, especially in the JJV corpus where the articles are on average short and the "protagonist" principal tie-up entity is highlighted at the outset by the Impact Line "wa." The protagonist entity usually announces the tie-up to the public, and in this sense, "has the action" throughout the remainder of the text. In short, when in doubt one should revert to the initial topic subject.

INVALID USES OF "WA"

Before turning to the Impact Line verbal element and finishing the prototype version-1 definition, the two types of occurrences below help illustrate further the legitimate uses of "wa" by showing what does not qualify as prototypical:

1. In the JJV test set, there are three instances in which the Impact Line topic marker is not preceded by an ENTITY but by a PERSON who is announcing a tie-up. The entity name is present as a modifier, e.g.,

日本開発銀行の高橋
Japan Development Bank's Takahashi

元・総裁 は
Hajime president <TM

Such instances are eliminated from consideration as a prototype because the initial "wa" is not preceded by a principal tie-up partner.

2. In one instance the initial “wa” marks a valid entity for extraction, however, it is not a principal tie-up partner; it is the PARENT of one of the principals.

(3) IMPACT LINE: OTHER REQUISITE ELEMENTS

As mentioned above under GRAMMATICAL FORCE, the JV application tracks tie-up relationships between two or more entities. And, it has already been demonstrated that the Impact Line topic marker is a reliable indicator (81% of the JJV test set) of at least one of those entities. The next question is: Does the prototypical Impact Line also contain the other elements required for instantiating a tie-up? That is: 1) Is the name of the other tie-up entity(ties) present in the Impact Line, and 2) is there any explicit indication that the arrangement between the two entities is in fact a tie-up relationship?

1) Remarkably, there are only seven instances -- over and above the previously cited 117 -- in which an Impact Line would otherwise be considered prototypical except that the other tie-up partner name(s) is not specified until later in the text. In other words, 81% of JJV test-set Impact Lines indicate clearly not only by virtue of the topic marker at least one tie-up entity, but also introduce the name of the other principal partner as well.

2) In order to confirm that any two or more entities present in the Impact Line are in a tie-up relationship, the Impact Line must state specifically that this is the case. The verbal elements at the end of the Impact Line are important to look at, therefore, in determining whether there is a tie-up or not.

Typically, Japanese text will stipulate “teikei," which is the most frequent term for tie-up, but will also use other phrases that are either synonymous or describe an arrangement or activity that presupposes a tie-up, such as:

結ぶことに合意
(agreed to join)

合併会社を設立するための契約
に調印した
(signed contract to establish JV
company)

研究開発契約を結んだと
発表した
(announced the formalization of an
R&D contract)

All of the previously judged 117 prototypical instance meet this standard, and not surprisingly, given the formulistic nature of the Impact Line, 96 out of those 117 (82%) employ the word “teikei.” (**Example** 7 later discusses an Impact Line in which “teikei” does not appear.)

(4) VERSION-1 REVIEW

Example 1:

```
東京海上火災      は  十七日
PN-Subject   <TM  Numeral+N
Tokyo Marine & Fire 17th

イギリス  の  大手総合保険
N       Prt      NP
English    big/gen'l/insur

会社  コマーシャル・ユニオン
 N          PN
comp. Commercial Union

社  (本社・ロンドン)    と
N     N      PN       Prt
comp. hqs. London     with
```

業務提携した　と　発表した。

VP　　Prt　　VP

business/tie-up/did　announcem't/did

Example 1 is reprised above to review the elements of a prototypical Impact Line. It must contain all the elements required by a valid tie-up. Therefore, the Impact line must state that there is a tie-up (or, was, in the case of dissolution) between at least two entities who are named; more if the partnership so stipulates.[4] Furthermore, at least one of the named tie-up entities -- the "protagonist" -- must be followed immediately by the topic marker "wa."

Version-1 Criteria:

- •Two Entities: Tokyo Marine & Fire and Commercial Union
- •"Wa-Designated Protagonist Tie-Up Entity: Tokyo Marine & Fire
- •Existence of Tie-Up Relationship: indicated by keyword 提携 "teikei"

At first glance this seems like an onerous burden for a prototypical structure to bear. But it is the discourse nature of Impact Lines in the JJV domain to be replete with pertinent information, much of it suitable for extraction. In view of the fact that the Impact Line introduces much data at the outset of an article, a more restrictive definition (version 2) requiring the Impact Line to contain additional extractable data items is presented below.

DEFINITION OF PROTOTYPICAL IMPACT LINE (VERSION 2)

The definition of version 2 requires the presence of two extractable data items in the Impact Line in addition to the minimum criteria of version 1. As the Impact Line in **Example** 1 above shows, a valid tie-up relationship exists between Tokyo Marine & Fire and Commercial Union. Moreover, the statement presents two additional pieces of information that are relevant for extraction: Commercial Union is an English company (NATIONALITY) and its headquarters is in London (ENTITY LOCATION). One is also told that Commercial Union is, indeed, a company (ENTITY TYPE), but this is considered less an item that is extracted discretely than one that follows automatically from the identification of the entity itself. This slot will be discussed later as a "default" fill.

The types of extractable data items that occur in the 117 prototypical Impact Lines are listed, with the SLOT NAME followed by instances of occurrence enclosed in parentheses: ENTITY LOCATION (79)*, INDUSTRY TYPE (88), PRODUCT/SERVICE (88), NATIONALITY (56)*, PERSON NAME (44)*, PERSON POSITION (40)*, PERSON ENTITY AFFILIATION (44)*, ALIAS (25), START TIME (12), END TIME (1), CHILD COMPANY (11), ECONOMIC ACTIVITY SITE (9), INVESTMENT (1), FACILITY NAME (1), FACILITY LOCATION (1), and JV COMPANY (1).

The *-marked slots indicate that when these particular data items appear in a JJV test-set article, they are more apt to appear in the Impact Line than in the remainder of the text. For example, ENTITY LOCATION information occurs in the Impact Line in 79 cases out of a total of 118 instantiations in the JJV test set, or 67% for the JJV test corpus; the percentages for PERSON NAME, PERSON ENTITY AFFILIATION, PERSON POSITION, AND NATIONALITY are 59%, 53%, 53%, and 44% respectively. There are,

[4] Two articles with 3 tie-up partners and one with 4 are included in the 117 prototypical cases.

moreover, orthographic consistencies in the textual presentation of certain information that should be noted: All but three of the 79 ENTITY LOCATION items are enclosed in parens; all but six for the ALIAS; and all of the PERSON NAME, POSITION, ENTITY AFFILIATION data.

Viewed another way, out of 117 version-1 prototypical Impact Lines, eight have no additional data items; 15 have just one; 27 have two; 19 have three; 17 have four; and 31 Impact Lines have five or more data items. In other words, if the version-2 definition of a prototypical Impact Line were to require the presence of two additional data elements, such as NATIONALITY and ENTITY LOCATION as in the case of **Example** 1 above, then there are 94 (117 minus the 23 that have less than two additional items) instances out of the 145 JJV test corpus that qualify, or 65% of the JJV test corpus. Viewed from either version of the Impact Line prototype, articles in the JJV test corpus possess at the outset a wealth of potential information for the extraction task -- 81% in its most lenient interpretation and 65% in its more restrictive.

Two Impact Line examples from the JJV test corpus are given below to highlight the requirements of the version-2 definition of the Impact Line prototype:

Example 6:

```
日立製作所          は 米国  の
 PN-Subj          <TM N+Prt(Adj)
Hitachi/manuf./place American
```

```
大手電算機メーカー
     NP
large/computer/maker
```

```
ヒューレット・パッカード社 (HP)
      PN
Hewlett Packard Co. (HP)
```

```
と    の    提携    を
Conj Prt   N     Prt
with     tie-up <DO marker
```

```
正式発表した。
    VP
formal/announcement/did
```

Translation:
Hitachi Manufacturing formally announced a tie-up with the large American computer maker, Hewlett Packard.

Version-2 Criteria

- •Two Entities: Hitachi Manufacturing and Hewlett Packard
- •"Protagonist" Tie-up Entity Marked by "wa": Hitachi Manufacturing
- •Tie-up Relationship: indicated by keyword 提携 "teikei"
- •Two Data Items: Nationality
  (American)
  Alias (HP)

Example 7:

```
アサヒビール    は  21日
   PN-Subj   <TM    N
Asahi/beer         21st
```

```
米国    の  生ビールメーカである
 N     Prt          NP
American  draft/beer/maker
```

```
アドルフ・クアーズ社
     PN
Adolph Coors Co.
```

```
(コロラド州)  の  ビール    を
    PN      Prt    N      Prt
(Colorado)        beer   <DO marker
```

```
国内    で  ライセンス生産し
Adj   Prt         VP
domestic  license/production/do
```

販売すること　になった　と
V+Nom(N)　VP　Prt
selling　was decided that

発表した。
VP
announcement/did

Translation:
On the 21st, Asahi Beer announced the decision that it will do the licensed production and selling of Adolph Coors' beer domestically; Adolph Coors (Colorado) is an American draft beer maker.

Version-2 Criteria

- Two Entities: Asahi Beer and Adolph Coors
- "Protagonist" Entity Marked by "wa": Asahi Beer
- Tie-up Relationship: indicated by phrases "produce" and "sell" that describe activities which presuppose tie-up
- Two Data Items (minimum):
  - Nationality (American)
  - Entity Location (Colorado)
- Additional Data Items Present:
  - Industry Type (Production)
  - Product/Service ("beer")
  - Industry Type (Sales)
  - Product/Service ("beer")
  - Economic Activity Agent (Asahi Beer)
- (Acceptable Additional Item:
  - Economic Activity Site (inference that "domestic" = Japan)

TEMPLATE DEFAULTS

Given the fact that the topic JJV sentence is stereotypical in both the amount of data contained (magnitude) and the way in which it is presented (Impact Line prototype), how this discourse structure might jump-start a system by providing top-level information which can be propagated throughout the template is examined next. One needs to discuss first, however, the notion of template "default" fills.

Default fills can be classified as either de jure, de facto, or logical. De jure defaults include the top-level or TEMPLATE OBJECT fills, such as the DOC-NR, DOC-DATE and DOC-SOURCE, whose slots are filled by SGML-tagged data items. They are, what one might call, "gimmes" by design and, therefore, are not incorporated in the scoring algorithm that measures system performance. The de facto and logical defaults need some explanation.

De facto defaults correspond to those set fills instantiated with a very high percentage of one type of data. Judging by actual systems' output and the patterns of certain answer-key template fills, no one will dispute that, in the end, data fell out of text into some set fills at a much higher frequency than was intuited originally when the template was being designed.[5] Below is a snapshot of high-percentage JJV test-set set fills. (The second figure represents percentages for 100 randomly selected development-set articles.)

[5] Some of the distinctions that were made at design time over the course of processing approximately 50 articles became blurred unavoidably as the fill rules evolved. Therefore, the initial random distribution between, e.g., the ENTITY TYPE set fills of COMPANY, GOVERNMENT, INDIVIDUAL, and OTHER became lopsided in favor of COMPANY.

SLOT NAME	FILL	TEST-SET%	DEV-SET%
TIE-UP STATUS	EXISTING	95%	91.50%
ENTITY TYPE	COMPANY	98.30%	96.60%
REL-ENT2-TO-E1	PARTNER	82.60%	84.50%
ENT REL STATUS	CURRENT	94.50%	95.50%

Given these percentages, how did the systems actually perform? Is there any indication that these de facto default fills were instantiated? The figures below seem to offer evidence for this. Every system evaluated on the TIPSTER JJV test corpus for MUC-5 showed substantially lower error rates for each of the above set fills versus their overall (All-Objects) error scores.

SYS-TEM	TIE-UP STAT-US	ENTI-TY TYPE	REL-2-TO-1	ER STAT-US	OVER-ALL ERROR
1	28	28	35	33	54
2	47	42	51	49	72
3	40	37	46	45	63
4	47	48	45	45	70
5	56	46	53	51	70
6	25	26	35	31	50

The descriptive analysis of the 12 templates mentioned above in METHODOLOGY shows a similarly distinctive trend in actual systems' output. The 12 templates were not randomly selected: All of them meet the version-1 definition for the Impact Line prototype, and only four do not meet the restrictive one; six articles are short -- six lines or less in length; one article specifies three principal tie-up partners in the Impact Line rather than the usual two; two articles contain multiple tie-ups rather than the usual (84% of JJV test corpus) one tie-up; one article specifically mentions the formation of a JV company in the Impact Line; two Impact Lines introduce a principal tie-up entity marked by the topic marker "wa" that is clausally modified by the name of its parent company; and one article's Impact Line marks two tie-up entities. In short, whenever a correct ENTITY was instantiated by any system, the above-mentioned default fills cascaded throughout the template, even if -- practically speaking -- the resulting fills indicated that a lone COMPANY was in a CURRENT PARTNER relationship with itself. The discussion of article 1528 below shows such an instance of this.

Other template fills can be regarded as logical defaults, or those that are a logical consequence of the template object-oriented design. If the keyword "teikei" confirms that there is a tie-up and its status is, as mentioned above EXISTING, then obviously the template has a tie-up event; i.e., a TIE-UP OBJECT must be instantiated to accommodate the extraction of such information as TIE-UP STATUS, ENTITY, etc. Similarly, if there is a tie-up event and two entities are in a relationship defined as PARTNER, then obviously there is an ENTITY RELATIONSHIP. If there is an INDUSTRY TYPE identified, there must be an ECONOMIC ACTIVITY OBJECT to accommodate the INDUSTRY OBJECT, which in turn accommodates the INDUSTRY TYPE. The template structure and other logical effects for inserting extracted data items into it will be outlined further below in the discussion of #1528.

THE COMBINED EFFECTS OF PROTOTYPICAL DISCOURSE AND THE DEFAULT MECHANISM

To illustrate the potential effects that stereotypical JJV discourse structure has on template fills and overall performance when the de facto defaults are considered as well, the example of article #1528 is submitted below.

1528 Impact Line:

```
資生堂   は   眼科用製薬会社
  PN <TM       PN
Shiseido ophthalmic/pharm./co.

千寿製薬              （本社大阪市
    PN                  N    PN
Senju Pharm'tical (hqs. Osaka

社長吉田祥二氏）
N    PN
pres./Yoshida/Shoji/Mr.)

整形外科用製薬会社   マルホ
     NP              PN
orthopedic/pharm./co. Maruho

（同、   山本秀夫氏）               と
  N        PN                     Conj
(ditto,Yamamoto/Hideo/Mr)         and

医療用医薬品          の   販売
     NP              Prt   N
medical/supplies          sales

で   提携した    と    発表した。。。
Prt    VP       Prt
  tie-up/did       announcement/did
```

Translation:
Shiseido announced that it had [concluded] a medical supplies sales tie-up with Senju Pharmaceutical (headquarters Osaka, Mr. Shoji Yoshida, president), a ophthalmic pharmaceutical company, and Maruho (ditto, Mr. Hideo Yamamoto), an orthopedic pharmaceutical company...(remainder omitted)

Number 1528 is a short six-line article with a version-2 prototypical Impact Line containing the following data items:

- Existence of Tie-up Relationship: indicated by keyword "teikei"
- "Protagonist" Tie-up Partner indicated by topic marker "wa": Shiseido
- Tie-up Partner: Senju Pharmaceutical
- Entity Location (specifically named): Osaka
- Person Name: Shoji Yoshida
- Person Position: President
- Entity Affiliation (info follows entity it describes): Senju
- Tie-up Partner: Maruho
- Entity Location (inferred from "ditto"): Osaka
- Person Name: Hideo Yamamoto
- Person Position: (unclear whether "ditto" indicates president)
- Entity Affiliation: Maruho
- Industry Type: Sales
- Product/Service String: "medical supplies"

Data items from remainder of text:

- Alternate Product/Service String for Sales
- Another Industry Type: Production
- Product/Service String for Production
- Alternate Product/Service String for Production
- Economic Activity Agents: Shiseido, Senju, Maruho
- Start Time for Production
- Revenue for Sales
- Start Time for Revenue
- Revenue Type
- Revenue Rate

Adding the logical and de facto default slots -- such as TIE-UP, TIE-UP STATUS, ENTITY TYPE, ENTITY RELATIONSHIP, REL-ENT2-TO-ENT1, ENTITY RELATIONSHIP STATUS, ECONOMIC ACTIVITY, etc., there are a total of 47 possible fills that are scored.

SYSTEM 1: MINIMUM CASE SCENARIO

Given the plethora of data items in the Impact Line and its prototypical structure, minimally a system should be able to identify and extract an ENTITY NAME (Shiseido) by the topic marker “wa” because this element of the Impact Line is the most consistent part of the prototype. Suppose, moreover, a system confirms the existence of a tie-up event (CONTENT) by identifying the keyword “teikei,” which is another consistent element of the Impact line prototype, and one other data item from the Impact Line such as the INDUSTRY TYPE SALES, which also has a keyword associated with it ”hanbai.” This system would have in effect identified and extracted three data items from the Impact Line. The default instantiations associated with the extraction of these items would be: TIE-UP STATUS (EXISTING), the named ENTITY (is a constituent of the TIE-UP), ENTITY TYPE (COMPANY), an ENTITY RELATIONSHIP, the named ENTITY (is a constituent of the ER), an ECONOMIC ACTIVITY (accommodates INDUSTRY), INDUSTRY (accommodates INDUSTRY TYPE), REL-ENT2-TO-ENT1 (PARTNER), and ENTITY RELATIONSHIP STATUS (CURRENT), for a total of 12 template fills.

This can also be viewed below schematically in template fashion. (The **bold** lettering indicates the three data items extracted from the Impact Line to highlight their place of insertion into the template and the embedding described above; *italicized print* indicates de facto default fills; plain text designates logical defaults; the <TEMPLATE OBJECT> de jure default fills are not scored except for CONTENT; and the numbers (1) - (12) represent the total correct fills.)

<TEMPLATE-1>:=
Doc Number: 1528
Doc Date: 900227
News Source: Nikkei Shimbun
Content: <TIE-UP-1> (1)
<TIE-UP-1>:=
Tie-up Status: Existing (2)
Entity: <ENTITY-1> (3)
Econ Activity:<ECON ACTIVITY-1> (4)
<ENTITY>:=
Entity Name: Shiseido (5)
Entity Type: Company (6)
ER:<ER-1>(7)
<ER-1>:=
Ent1: <ENTITY-1> (8)
Rel-Ent1-To-Ent2: Partner (9)
Status: Current (10)
<ECON ACTIVITY-1>:=
Industry: <INDUSTRY-1> (11)
<INDUSTRY-1>:=
Industry Type:Sales (12)

To review the logic outlined above: An entity name is correctly identified by the topic-marker heuristic; in order to place the name within the template, an ENTITY OBJECT must be generated to accommodate it; this is accomplished through the generation of a TIE-UP OBJECT which, in turn, is generated by the CONTENT pointer; CONTENT is confirmed by the keyword “teikei;” the third data item “sales” can be inserted into the template once an ECON ACTIVITY OBJECT is generated in order to accommodate the INDUSTRY OBJECT needed to instantiate the INDUSTRY TYPE data; if a named ENTITY is inserted as above, it, by definition, must be a constituent part -- or principal partner -- of a TIE-UP, and also, by definition, must be in an ENTITY RELATIONSHIP with another entity (not identified here); the rest of the slots are de facto default fills.

The results of identifying and extracting successfully three data items from the Impact Line would be as follows:

•12 slots are filled out of a possible total of 47
•All 12 are correct
•Recall = 26
•Precision = 100
•Error = 74
•Undergeneration = 74

This means that what the system did capture, it did so accurately; and it did so through the identification of only a small percentage of the data items available to it in the Impact Line. Through the "default" mechanism, three discrete elements proliferated into a template with 12 correct fills.

SYSTEM 2: BETTER CASE SCENARIO

Suppose, however, another system, System 2, extracts successfully the same three data items as System 1 and, in addition, identifies other Impact Line information such as ENTITY LOCATION (Osaka), PERSON NAME (Shoji Yoshida), PERSON POSITION (President), ENTITY AFFILIATION (Shiseido), and another named ENTITY (Senju). System 2, moreover, successfully recognizes a START TIME which appears in text after the Impact Line. Finally, this system incorrectly extracts a second INDUSTRY TYPE (RESEARCH rather than PRODUCTION), and lists only two ECON ACTIVITY AGENTS (Shiseido and Senju) rather than three (Shiseido, Senju, and Maruho) because it failed to identify the third entity name in the Impact Line. System 2, in short, has done a better job than System 1 in making use of the top-level Impact Line data available to it. However, it still misses several Impact Line items and misidentifies (undergenerates) two others, but coupled with the instantiation of the same defaults outlined in the schematic above the results would look more impressive:

•Out of 47 total possible scored slots, 29 are filled; 26 correctly.
•Recall = 55
•Precision = 90
•Error = 46
•Undergeneration = 40

SYSTEM 3: BETTER STILL

Finally, suppose yet another system, System 3, does an even more thorough job of extracting data from the Impact Line. In addition to what System 2 recognizes, this system identifies the third entity (Maruho), a second PERSON (Hideo Yamamoto) with ENTITY AFFILIATION (Maruho) and POSITION (infers "President" from "ditto" which is scored as acceptable), and the PRODUCT/SERVICE string associated with SALES. Like System 2 above, System 3 recognizes a START TIME from the body of the text and misidentifies a second INDUSTRY TYPE as RESEARCH. Since this system has managed to extract every piece of Impact Line information and insert it into the template along with the default fills, not surprisingly its results would look impressive indeed.

•Out of 47 possible scored slots, 38 are filled; 37 correctly.
•Recall = 80
•Precision = 99
•Error = 20
•Undergeneration = 19

CONCLUSION

This paper has shown that JJV articles possess a stereotypical pattern of introducing much significant information amenable to the data extraction task. This stereotypical pattern is embodied in what has been outlined here as the

Impact Line prototype. Furthermore, the "mining" of the Impact Line to a minimal degree by extracting the topic marker-designated ENTITY is, one could say, a little that goes a long way. This is due in large part to that ENTITY's strategic place in the template and the way in which default fills associated with it are propagated throughout the template. Hence, higher scores result for JJV than EJV.

A system, such as System 3 above, that takes full advantage of the Impact Line prototype and the plethora of information available therein can maximize its capability and show a quantum leap in statistical performance. Obviously, the formulation of a complete JJV discourse structure would raise performance to another level.

Discourse analysis alone, however, will not resolve all the problems endemic to Japanese, such as ellipsis. If the formulistic nature of Japanese discourse in the JJV domain is a boon to data extraction, then its penchant for omitting sentence topics altogether is a potential minefield. Discrete data items that have been easily identified at the outset need to be correctly referenced to other activities that follow or the resulting template fills will paint a totally misleading picture as to who is doing what to whom. This paper has discussed a heuristic for topic-marker substitution that might help in this regard, but it is only a small part of the equation for making Japanese more explicit.