THE MURASAKI PROJECT: MULTILINGUAL NATURAL LANGUAGE UNDERSTANDING

Chinatsu Aone, Hatte Blejer, Sharon Flank, Douglas McKee, Sandy Shinn

Systems Research and Applications (SRA)
2000 15th Street North
Arlington, VA 22201

ABSTRACT

This paper describes a multilingual data extraction system under development for the Department of Defense (DoD). The system, called Murasaki, processes Spanish and Japanese newspaper articles reporting AIDS disease statistics. Key to Murasaki's design is its language-independent and domain-independent architecture. The system consists of shared processing modules across the three languages it currently handles (English, Japanese, and Spanish), shared general and domain-specific knowledge bases, and separate data modules for language-specific knowledge such as grammars, lexicons, morphological data and discourse data. This data-driven architecture is crucial to the success of Murasaki as a language-independent system; extending Murasaki to additional languages can be done for the most part merely by adding new data. Some of the data can be added with user-friendly tools, others by exploiting existing on-line data or by deriving relevant data from corpora.

1. INTRODUCTION

Project Murasaki is a 30-month project for DoD to design and develop a data extraction prototype, operative in Spanish and Japanese and extensible to other languages. Using SRA's core natural language processing (NLP) software, SOLOMON, Project Murasaki extracts information from newspaper articles and TV transcripts in Japanese and from newspaper articles from a variety of Spanish-speaking countries. The topic of the articles and transcripts is the disease AIDS. The extracted information – some in a canonical form and some as it appears in the input texts – is stored in an object-oriented database schema implemented in a recently released multilingual version of the Sybase RDBMS.

Project Murasaki has been under development since October 1990 and will be delivered to DoD in June 1993. The goal of the project was to extend SOLOMON's data extraction capabilities, hitherto used for English texts, to Spanish and Japanese. It was explicitly requested that Murasaki be as language-independent and domain-independent as possible and be extensible to additional languages and domains ultimately.

SOLOMON reflects six years of development. From its inception, language and domain independence have been deliberate design goals. Murasaki was our first extensive use of SOLOMON for languages other than English and thus the first testing-ground for its claimed language independence. SOLOMON had been used and continues to be used across a variety of domains over the past six years. In the MUC-4 conference, SRA demonstrated a single system extracting information about Latin American terrorism from newspaper articles in all three languages, using Spanish and Japanese data modules developed for Murasaki and terrorism vocabulary in Spanish and Japanese acquired in the two weeks prior to the demonstration (cf. [1, 2]).

SOLOMON's architecture did not change significantly during the course of Murasaki. For the most part, its claim to language independence was borne out. Below, we will discuss how we have extended it to increase its language independence.

2. UNIQUE FEATURES OF MURASAKI

2.1. Modular Architecture

Murasaki is composed of shared processing modules across the three languages supported by separate data modules, as shown in Figure 1. Murasaki has six processing modules: PREPROCESSING, SYNTAX, SEMANTICS, DISCOURSE, PRAGMATICS, and EXTRACT. Each of these modules has associated data. For example:

PREPROCESSING:	lexicons, patterns, morphological data
SYNTAX:	grammars
SEMANTICS:	knowledge bases
DISCOURSE:	discourse knowledge sources
PRAGMATICS:	inference rules
EXTRACT:	extract data

Modularity is crucial to the *reusability* and *extensibility* of Murasaki. It facilitates, on the one hand, reuse of parts of Murasaki and on the other hand replacement of parts of the system.

We have been able to reuse portions of SOLOMON in the past, and expect to be able to pull modules out of Murasaki and use them separately as warranted in the future. For instance, PREPROCESSING could be used in isolation in multilingual information retrieval applications.

Conversely, modules – both processing and data modules –

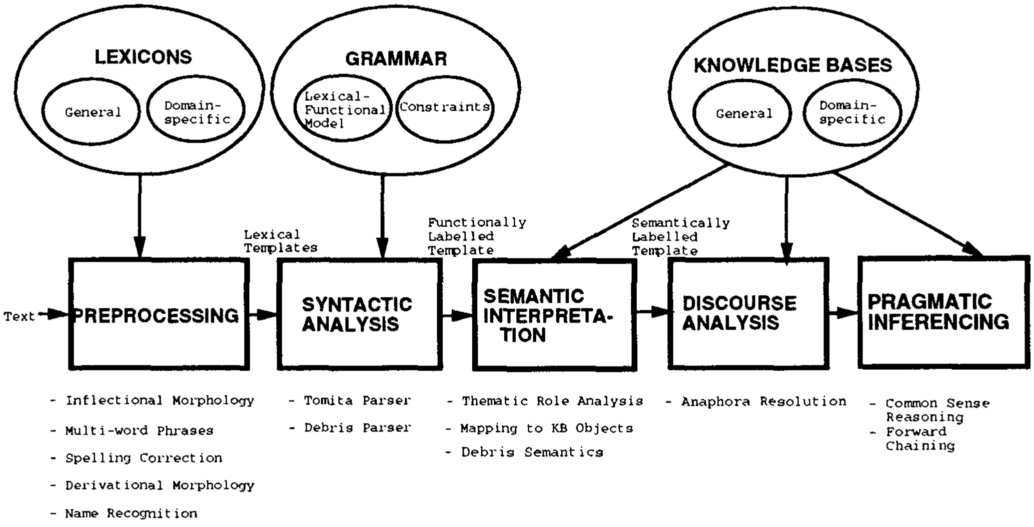

Figure 1: Murasaki Architecture

can be replaced as technology improves or in order to port to a new language or new domain. In order to port Murasaki to MUC-4 Latin American terrorism domain, we replaced the Japanese and Spanish AIDS domain lexicons with Japanese and Spanish terrorism domain lexicons, resulting in a system which understood Spanish and Japanese newspaper articles on terrorism in a matter of weeks. Since the terrorism domain knowledge bases (KB's) were developed for English for MUC-4 already, and since the KB's can be shared across languages, there was no need to change or add KB's in this case.

In addition to plugging in new data modules, we have successfully replaced single processing modules, (separately) at various times, such as PREPROCESSING, SEMANTICS, DISCOURSE and EXTRACT without changes to the other modules. In addition, we added an entirely new module PRAGMATICS in the past year. In no case were extensive changes to other parts of the system required.

Finally, in developing NLP systems it is crucial to be able to isolate the source of system errors during development and testing. While black-box testing can indicate how the system is performing overall, only glass-box (module-level) testing can focus on the source of errors in such a way as to aid the developers. Murasaki's modular architecture has facilitated such glass-box testing.

2.2. Data-driven Architecture

Each Murasaki processing module is data-driven. Data modules are specific to the language, domain, or application. Keeping the data modules separate is an essential factor in the system's success as a multilingual system. We have been able to isolate the majority of the language-specific knowledge to the data modules associated with PREPROCESSING (i.e. lexicons, patterns, morphological data) and SYNTAX (i.e. grammars). SEMANTICS is entirely language-independent, and DISCOURSE isolates the small amount of language-specific information to the discourse data module (i.e. discourse knowledge sources).

Thus, in order to port to a new language, the following subset of the data modules are necessary:

PREPROCESSING:	lexicons, patterns, morphological data
SYNTAX:	grammars
DISCOURSE:	discourse knowledge sources

To facilitate data acquisition in multiple languages, we have been developing language-independent automatic data acquisition algorithms [3, 4]. Also, in order to improve the quality of grammars, we have adapted a grammar evaluation tool (PARSEVAL) to evaluate the performance of our Spanish and Japanese grammars on texts bracketed by the Penn Treebank bracketing tool.

3. MULTILINGUAL MODULES

In this section, we discuss what we have done to the processing modules in Murasaki in order for them to handle multilingual input.

3.1. Preprocessing

Murasaki replaced its original morphological analyzer with a multilingual morphological analyzer. The new analyzer consists of a morphological processing engine and morphological data for each language, as shown in Figure 2. In order to add a new language, one only has to add morphological data for

```
VSTEM TENUMPERS ER-VERB PRETERIT
           com er
           yo         com   !'i
           t!'u       com   iste
           ella       com   i!'o
           nosotros   com   imos
           vosotros   com   isteis
           ellos      com   ieron
```

Figure 2: An Example of Spanish Morphological Data

the language. This approach is especially useful for highly inflected languages like Spanish and Japanese.

After morphological analysis, pattern matching is performed to recognize multi-word phrases like numbers, date, personal names, organization names, and so on. Although specific patterns to recognize, for example, Japanese and Spanish personal names are different, the same pattern matching engine and pattern specification language are used for all the languages. Examples of phrases recognized by Spanish patterns are shown in Figure 3, using SGML markers.

3.2. Syntax

In order to make the syntax module language-independent, Murasaki extended the Tomita parsing algorithm [5] to deal with ambiguous word boundaries in Japanese. These boundaries are problematic because there is no space between words in a Japanese sentence, and it can be segmented in more than one way. Our implementation of the algorithm was originally token-based, where a token was a word or phrase. The extended algorithm is character-based, and now allows variable length input.

The same X-bar-based grammar specification language for English has been used to write the Spanish and Japanese grammars. In addition, all the grammars call the same constraint functions to check syntactic subcategorization and semantic type restrictions during parsing. Skeletal rules can be provided for a new language to start with especially when the new language is structurally similar to the languages of the existing grammars (e.g. Portuguese). In fact, much of the Spanish grammar was derived from the English grammar. A few basic X-bar rules for head-initial (e.g. English and Spanish) and head-final languages (e.g. Japanese) are shown in Figure 4.[1]

```
##054 09ago89 Excelsior-Jalapa palabras 218
     El nu'mero de casos de sida en la entidad
aumento' a 326, con los 15 detectados durante
<time>el mes de julio</time>, aseguro' hoy
<name>el doctor Jose' Rodri'guez Domi'nguez</name>,
jefe de <org>los Servicios Coordinados de Salud
Pu'blica</org> en el estado.
```

Figure 3: Spanish Text with Pattern Examples

Head-Initial	Head-Final
V1 → *V NP	V1 → NP *V
P1 → *P NP	P2 → NP PARTP
GMOD → SUBCONJP S	GMOD → S SUBCONJP
N4 → N4 GMOD	N4 → GMOD N4
N4 → N4 NCOMPS	N4 → NCOMPS N4

Figure 4: Examples of Basic X-Bar Rules

The output of the parser is a structure called a functionally labeled template (FLT), which is similar to LFG's f-structure. The FLT specification language is language-independent. It uses grammatical functions like subject, object, etc. as registers, but no language-specific information such as precedence is present at this level. Thus, while Spanish texts often use inversion as in "... y en total se han registrado cuarenta y siete casos con treinta víctimas", it is not the case with English, e.g. ".. and in total 47 cases with 30 victims were recorded." However, such differences are normalized in FLT's.

The FLT specification has been extended and tested to cover phenomena in three languages as Spanish and Japanese grammars are developed. It must be general and expressive enough to cover linguistic phenomena in multiple languages because the semantic interpretation expects its input in any language to follow this specification. For example, quantity phrases (QP's) in any languages now have a *unit* register in the FLT. In English and Spanish, measure units in measure phrases fill the unit values (e.g. "3 *pints* of blood", "62 *por ciento* de las personas con sida"). In Japanese, so-called classifiers fill in the unit values. Classifiers are used to count any objects in Japanese, including discrete objects such as people and companies. Such unit information is sometimes important for semantic disambiguation.

Once broad-coverage grammars are developed, they can be used to process texts in any domain (unless the domain uses a sublanguage) without much modification, since linguistic structures of languages do not change from domain to domain. The only difference between domains is the weights on each rule because the frequency of using particular rules changes, just as the frequency of particular words changes in different domains. Thus, the same Spanish and Japanese grammars developed for the AIDS domain were used to process Spanish and Japanese texts in the MUC-4 (terrorism) domain with a

[1] GMOD stands for General MODifier, NCOMPS for Noun COMPlementS, and PARTP for PARTicle Phrase. The Arabic numerals indicate bar levels.

few additional rules.

3.3. Semantics

Murasaki has a single semantics processing module for all the languages. It takes as input output of the syntax module in the FLT format. Thus, so long as the input to the semantics processing module conforms to the FLT specification, Murasaki can use different grammars or parsers. The semantics processing module uses core and domain-specific knowledge bases (KB's) common to all languages to perform semantic disambiguation and inference, and outputs language-independent KB objects.

In moving from English to multiple languages, English specific information was moved from the KB's to the lexicons. Our KB's originally encoded both semantic and English-specific syntactic information (e.g. subcategorization information). We moved the language-specific syntactic information from the KB's to English lexicons, and left the language-independent semantic type information in the KB's.

In addition, the semantics processing module itself has become more data-driven to be more language-independent. For example, interpretation of the complements of nominalized verbs among three languages became language-independent by classifying pre/postpositions of these languages into common semantic classes. Thus, AGENT roles of nominalized verbs are typically expressed by AGENT markers (e.g. "by", "por", "niyoru") of given languages, and THEME roles by NEUTRAL markers (e.g. "of", "de", "no").

	AGENT	THEME
Eg	investigation *by* WHO	transmission *of* AIDS
Sp	investigaci6n *por* WHO	transmisi6n *de* SIDA
Jp	WHO-*niyoru* chousa	AIDS-*no* kansen

A more general, data-driven approach has been also taken for semantic disambiguation necessary to interpret pre/postpositional phrases, compound nouns, appositives etc. which are common for all the languages. In all cases of semantic disambiguation, the same knowledge-based strategy is used to determine the most plausible relations between two semantic objects. For example, for noun phrases like "AIDS/cancer patients", "afectados de SIDA" (Spanish), or "AIDS kanja" (Japanese), a relation **Has-Disease** is chosen from the KB's for the two nouns. Semantics of ambiguous pre/postpositions (e.g. "in", "en") is determined in a similar way. For sentences in (1) below, a relation **Location** is chosen, while for those in (2) a relation **Time** is chosen.

(1) a. 500 men were infected with AIDS in China.
b. En China se han infectado 500 hombres con SIDA.
(2) a. 500 men were infected with AIDS in March.
b. En marzo se han infectado 500 hombres con SIDA.

```
(WAKARU
        ((CATEGORY . V)
         (INFLECTION-CLASS . CR)
         (GLOSS . UNDERSTAND)
         (PREDICATE #UNDERSTAND#)
         (SITUATION-TYPE INVERSE-STATE)
         (IDIOSYNCRASIES
           (GOAL (MAPPING (LITERAL ''NI'')
                          (SURFACE SUBJECT))))
         (TRANSCRIPTION . WAKARU)))
```

Figure 5: A Lexical Entry for "wakaru"

The Murasaki semantics module uses four basic language-independent predicate-argument mapping rules called *situation types*, which map syntactic arguments of verbs in FLT's (e.g. subject, object, etc.) to thematic roles of verb predicates in the KB's (e.g. agent, theme, etc.), as shown in Table 1. Such mapping rules, along with any idiosyncratic mapping information, for each verb are derived from corpora automatically (see [4] for more detail).

For example, the English word "understand" uses what we call INVERSE-STATE mapping, where the subject maps to GOAL and and the object THEME of the predicate #UNDERSTAND#. The Japanese semantic equivalent "wakaru" also uses the INVERSE-STATE mapping. However, the language-specific idiosyncratic information that the GOAL role can be also specified by a particle "ni" is stated in the lexicon. As shown in Figure 5, the lexical entry for "wakaru" has pointers to its semantic predicate in a KB (i.e. #UNDERSTAND#) and its mapping rule (i.e. INVERSE-STATE), and specifies its word-specific idiosyncrasy information about "ni" in addition.

3.4. Discourse

The Murasaki discourse module needed the most work to be language-independent. A discourse module is generally least developed in any NLP system, and our system was no exception. In addition, some part of the module was designed to be English specific. For example, since grammatical genders and natural genders usually coincide in English, the original discourse module paid attention only to natural genders. However, in Spanish, grammatical genders of an anaphor and its antecedent, not the natural genders, must be compatible for them to co-refer. For example, the third person feminine pronoun "la" in the following sentence refers to "la transmisi6n", which is not a semantic object with a female gender: "En otras entidades como Baja California y Veracruz la transmisi6n en este grupo es 1.2 veces mayor que la que ocurre a nivel nacional."

Moreover, different languages have different types of anaphora (e.g. zero pronouns in Spanish and Japanese). In

Situation Types		English/Spanish Mapping	Japanese Mapping
CAUSED-PROCESS	AGENT	(SURFACE SUBJECT)	(SURFACE SUBJECT)
	THEME	(SURFACE OBJECT)	(SURFACE OBJECT)
PROCESS-OR-STATE	THEME	(SURFACE SUBJECT)	(SURFACE SUBJECT)
AGENTIVE-ACTION	AGENT	(SURFACE SUBJECT)	(SURFACE SUBJECT)
INVERSE-STATE	GOAL	(SURFACE SUBJECT)	(SURFACE SUBJECT)
	THEME	(SURFACE OBJECT)	(SURFACE OBJECT) (PARTICLE "GA")

Table 1: Predicate-Argument Mapping Rules (Situation Types)

addition, languages differ in the distribution patterns of each type of anaphora (e.g. the antecedent of a Japanese anaphor "uchi" is found in the adjacent *discourse clause*). Furthermore, constraints on the antecedents differ from language to language (e.g. a Japanese third person masculine pronoun "kare" must refer to a male person, but not Spanish third person masculine pronouns).

We achieved the multilingual capability of the discourse module by dividing the anaphora resolution process into multiple knowledge sources and using subsets of the knowledge sources to handle different discourse phenomena (cf. [6]). Both the discourse knowledge sources and discourse phenomena are represented as objects in the KB's. Thus, the discourse processing module called *Resolution Engine* has become strictly data-driven (cf. Figure 6).

The discourse knowledge source KB consists of *generators* (i.e. various ways to generate antecedent hypotheses), *filters* (e.g. syntactic number filter, syntactic gender filter, semantic amount filter, semantic gender filter, semantic type filter, etc.), and *orderers* (e.g. focus orderer, recency orderer, etc.). Language-independence of the knowledge sources has been achieved by dividing each knowledge source into language-specific data and language-independent processing functions. For example, the semantic gender filter has associated data for English and Japanese, which specifies constraints on genders of semantic objects imposed by certain pronouns (e.g. English "he" cannot refer to semantic objects with female gender like "girl"). As explained above, Spanish does not use the semantic gender filter but uses the syntactic gender filter.

Finally, we wanted to be able to evaluate the performance of the Murasaki discourse module so that we can train it and maximize its performance in different languages and domains. Our architecture allows anaphora resolution performance to be evaluated and trained. We use corpora tagged with discourse relations, as shown in Figure 7, for such evaluation.

4. CONCLUSION

We have described a multilingual system, Murasaki, focusing on specifics of its language-independent architecture and describing how language-specific data is integrated with general processing modules. While this architecture is currently operating for data extraction from Japanese, Spanish, and English texts, it has been designed to be extended to additional languages in the future. Murasaki also has associated multilingual data acquisition tools and algorithms, which have been used to extend its data modules. In addition, we have developed preliminary multilingual training and evaluation tools for the syntax and discourse modules of Murasaki.

Planned future enhancements include addition of new data modules (e.g. multilingual "WordNets"), extension of the Spanish and Japanese data sources to new domains, and improved multilingual tools for automatic data acquisition from corpora. We would also like to extend the system to a new, typologically different language such as Arabic in order to further test and refine its language independence.

References

1. Aone, C., McKee, D., Shinn, S., and Blejer, H., "SRA: Description of the SOLOMON System as Used for MUC-4," in *Proceedings of Fourth Message Understanding Conference (MUC-4)*, Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1992.
2. Aone, C., McKee, D., Shinn, S., and Blejer, H., "SRA SOLOMON: MUC-4 Test Results and Analysis," in *Proceedings of Fourth Message Understanding Conference (MUC-4)*, Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1992.
3. McKee, D., and Maloney, J., "Using Statistics Gained From Corpora in a Knowledge-Based NLP System," in *Proceedings of The AAAI Workshop on Statistically-Based NLP Techniques*, 1992.
4. Aone, C., and McKee, D., "Three-Level Knowledge Representation of Predicate-Argument Mapping for Multilingual Lexicons," in *AAAI Spring Symposium Working Notes on "Building Lexicons for Machine Translation"*, 1993.
5. Tomita, M., "An Efficient Context-free Parsing Algorithm for Natural Language," in *Proceedings of IJCAI*, 1985.
6. Aone, C., and McKee, D., "Language-Independent Anaphora Resolution System for Understanding Multilingual Texts," to appear in *Proceedings of 31st Annual Meeting of the ACL"*, 1993.

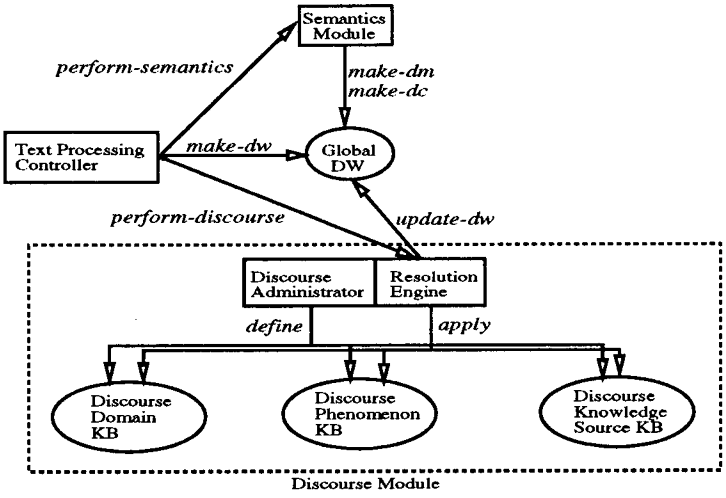

Figure 6: Discourse Architecture

La Comisio'n de Te'cnicos del SIDA informo' ayer de que existen <DM ID=2000>196 enfermos de <DM ID=2001>SIDA</DM></DM> en la Comunidad Valenciana. De <DM ID=2002 Type=PRO Ref=2000>ellos</DM>, 147 corresponden a Valencia; 34, a Alicante; y 15, a Castello'n. Mayoritariamente <DM ID=2003 Type=DNP Ref=2001>la enfermedad</DM> afecta a <DM ID=2004 Type=GEN>los hombres</DM>, con 158 casos. Entre <DM ID=2005 Type=DNP Ref=2000>los afectados</DM> se encuentran nueve nin'os menores de 13 an'os.

Figure 7: Discourse Tagged Corpora