The SYSTRAN Machine Translation System

ARPA MT Workshop
Vienna, Virginia , March 16-18 1994

Elke Lange, Senior Linguist
SYSTRAN Translation Systems, Inc

SYSTRAN Components

Systems Control Software

Linguistic Software

Peripheral Software

Input/Output Filters
(Text Preprocessors)

Development/Diagnostic Tools

Parse Diagnosis (PDIAG)
Translation Output Comparison
Concordances

Languages Translated by SYSTRAN

Russian-English

English-Source

English	→ French	(1)
	→ German	(1)
	→ Spanish	(1)
	→ Italian	(1)
	→ Portuguese	(1)
	→ Dutch	(2)
	→ Danish	(2)
	→ Swedish	(2)
	→ Norwegian	(2)
	→ Finnish	(2)
	→ Arabic	(1)
	→ Japanese	(3)
	→ Russian	(2)
	→ Greek	(3)
	→ Korean	(3)

Asian Source Languages

Japanese	→ English	(2)
Korean	→ English	(3)
Chinese	→ English	(4)

German-Source

German	→ English	(1)
	→ French	(2)
	→ Italian	(3)
	→ Spanish	(3)

Romance-Source

French	→ English	(1)
	→ German	(2)
Spanish	→ English	(2)
Italian	→ English	(3)
	→ French	(3)
Portuguese	→ English	(3)

Other:

Arabic	→ English	(4)

Notes: 1. operational system 2. "young system" 3. pilot system 4. planned for 1995

Increasing Modularity

Single Source → Single Target
(Russian-English "language pair")

↓

Single Source → Multiple Target
(English-Mltgt, German-Mltgt)

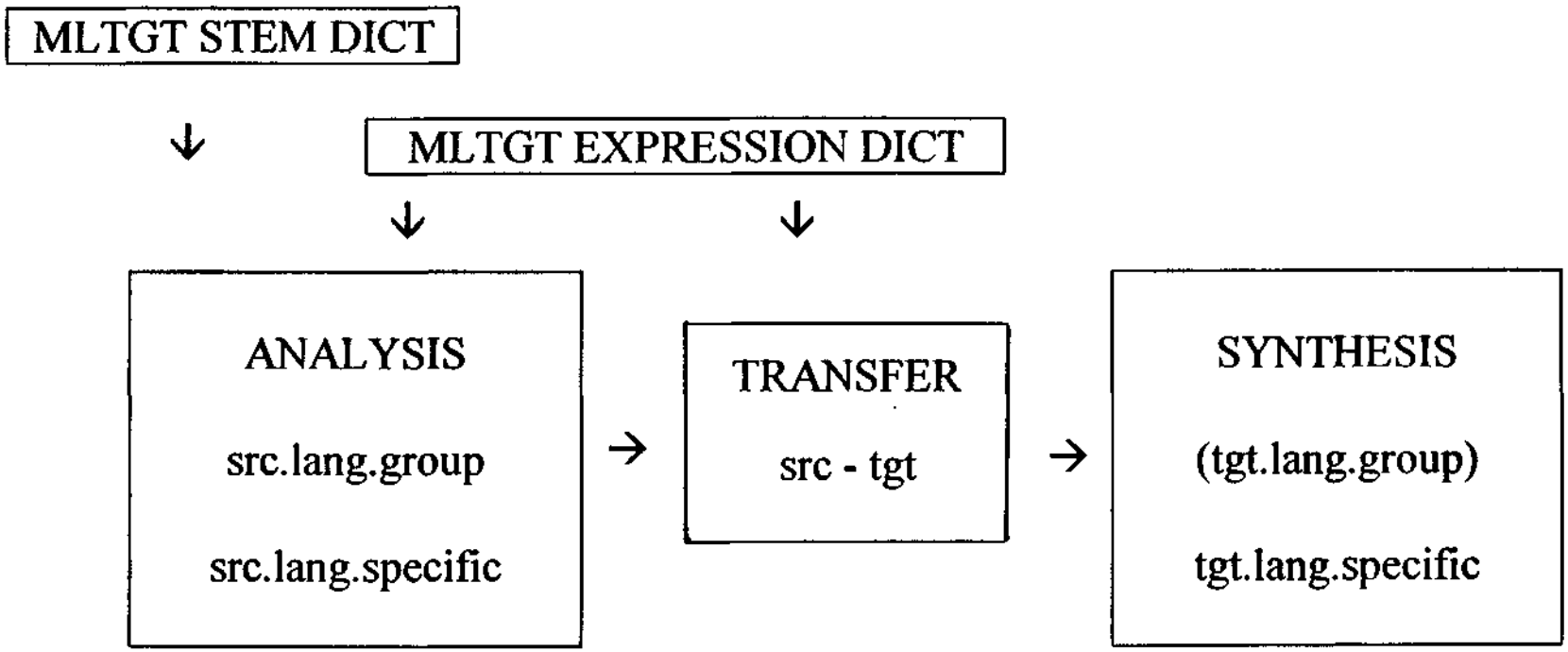
↓

Multiple Source → Multiple Target
(Romance Source)

↓

future
?
(Single Language Source/Target)
(already shrinking transfer programs;
single language source/target dictionaries)

Linguistic components



Dictionaries
(total 1.2 million entries)

Stem Dict.

SRC Lang. information
- single words
- stem & morphology code
- homograph identifier
- cross-referencing
- syntactic codes
- semantic codes
- domain codes

Expression Dict.

SRC Lang. rules
- contiguous (e.g. noun phrases)
- discontinuous expressions
- over 400 macros for rules
- flexible syntax
- most set target meanings
- some do source disambiguation
- reference to stem dictionary

TGT Lang. information
- meaning & morphology code
- any / all target lang.
- subject field distinctions
- customer specific meanings
- tgt lang. variant

Source Language Analysis

(1) Word level:
- morphology
- not-found-word analysis
- early disambiguation
- homograph resolution

(2) Clause level:
- clause boundaries
- clause types
- embedding levels

(3) Surface Syntax:
- basic syntactic relationships
- enumeration, extension of basic relationships
- subject / predicate , review of noun phrase function
- preposition analysis

(4) Deep Syntax:
- semantic relationships

(5) Review of entire Parse:
- filter for apparent analysis errors

(5) Special Lexicals:
- source language lexical disambiguation

Language Pair Transfer

Transfer Programs
- lexical transfer
- preposition translation
- structural transfer
- rearrangement of word order

Target Language Synthesis

- syntactic rules
- morphology

Characteristics of the SYSTRAN Parser

Entire Text
- incl. incomplete / ungrammatical sentences
- some text segmentation (determine sentence end)
- sentence segmentation (logical break)
- word segmentation (Japanese word boundaries)

Sentence by Sentence
but:
- buffer area for information from previous sentence(s)
- "memory feature" allows polling of information over entire paragraph

10 - 12 Major Passes

Deterministic
but:
- next pass may detect error and correct
- uncertain analysis decisions are flagged
- Analysis Filter Program
  (error information used by transfer and synthesis)
  (recursion to previous pass for alternate analysis path possible in current architecture)

Detailed Rule Base
- no theoretical constraints
- sound basic rules
- unlimited expansion
- based on large databases of live text
- years of detailed analysis

Language Independent
- little or no reference to actual words
- extensive use of dictionary codes
- same program flow
- same algorithms

Abstract Representation
- same abstract representation for all languages