STILVEN
a Veneto-English - and viceversa - Automatic Translator

Rodolfo Delmonte
Department of Language Sciences
Ca' Bembo, Dorsoduro 1075
30123 - VENEZIA (Italy)

OUTLINE

- Translation Modes
- Work Done last year
- Problems related to Veneto
- Solutions proposed and technology chosen
- Conclusions

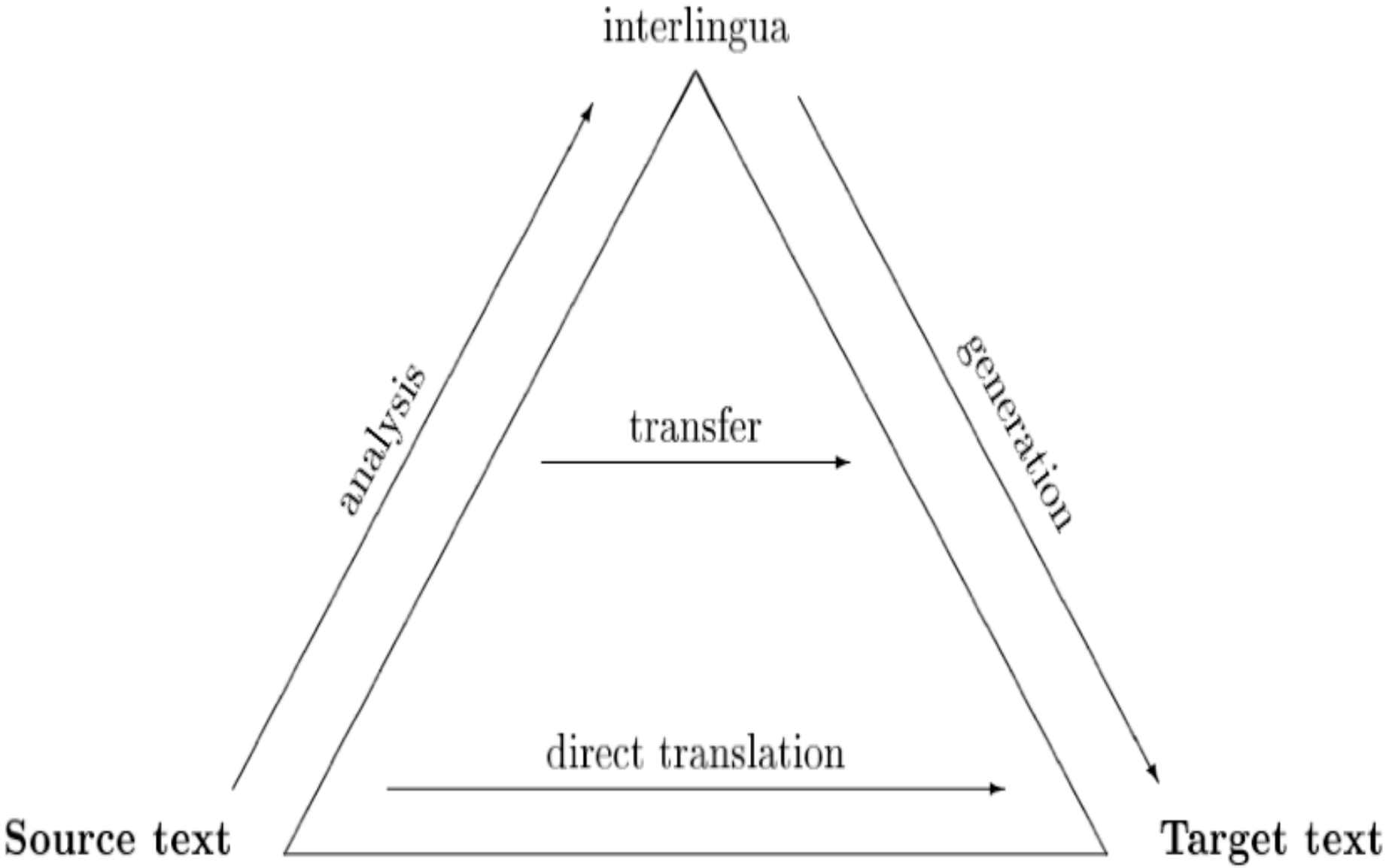


Figure 3.1: Vauquois' pyramid

Translation Modes

- (Data Driven) Statistical Machine Translation
  - 100million word parallel corpus (optimal)
- Direct Translation
  - No linguistic intermediate representation
- Indirect Translation
  - Interlingua
    - Semantic/Pragmatic linguistic representation
  - Transfer
    - Linguistic representation Lexical/Morphological/Syntactic

Data Driven MT

- Parallel Aligned Texts
- Construct a so-called translation model, giving for each possible sentence, the probability that this sentence may be translated to a specific target language string
- The best translation is the target language string with the highest probability
- Lots of constraints are introduced to restrict the search when the model becomes too big

Data Driven MT

- Often words are not translated one to one, but one word in one language is linked to two, three or more words in the other
- This is called the fertility of the word
- At other times some words, usually grammatical words like particles and prepositions, do not correspond to any words in the translation
- This is seen as a fertility of zero degree

Data Driven MT

- Some of the constraints may regard:
  - The length of the source and target language sentences
  - The position of both the source and the target word in their sentences
  - The distance intervening between the two positions which is fixed to a given threshold
  - The fertility of the source word
  - The position of the other target words that partake in translating the source word

Data Driven MT

- The task of translation is then, given the translation model, to find the string which is the most likely translation.
- That is an estimation maximization task
- But instead of maximizing the formula Pr(t|s), i.e. given the source sentence s, which is the most probable targer translation?
- On the contrary they try to solve the equivalent Pr(t)Pr(s|t), i.e. the target sentence which is the most likely to have been the source, if the source sentence had been the target (the one that the native speaker had in mind when he translated in the target language)

Data Driven MT

- In fact, in this way, we reduce the problem of MT to that of finding the best grammatical (i.e. actually occurring, or observed) sequence of words (sentence/phrases as would be in the case of MOSES) in the target language given a certain tranlastional pair
- This is a bigram or trigram statistical model of the target language, i.e. for each string of two or three English words finding the probability that words with given tags occurr in this order
- Statistical approaches are extremely vulnerable to the data sparsity problem: one occurrence is not sufficient because hapax legomena will be automatically discarded by the MT system

Data Driven MT

- Morphological analysis may help in reducing the wordforms
- Improving on statistical results is hard, it is always a take-it-or-leave-it
- Attempts at inserting linguistics in the MT process has been disappointing so far

- Other cases of DDMT are Example Based MT systems which use direct match between the corpus and the new sentence to translate

Direct Systems

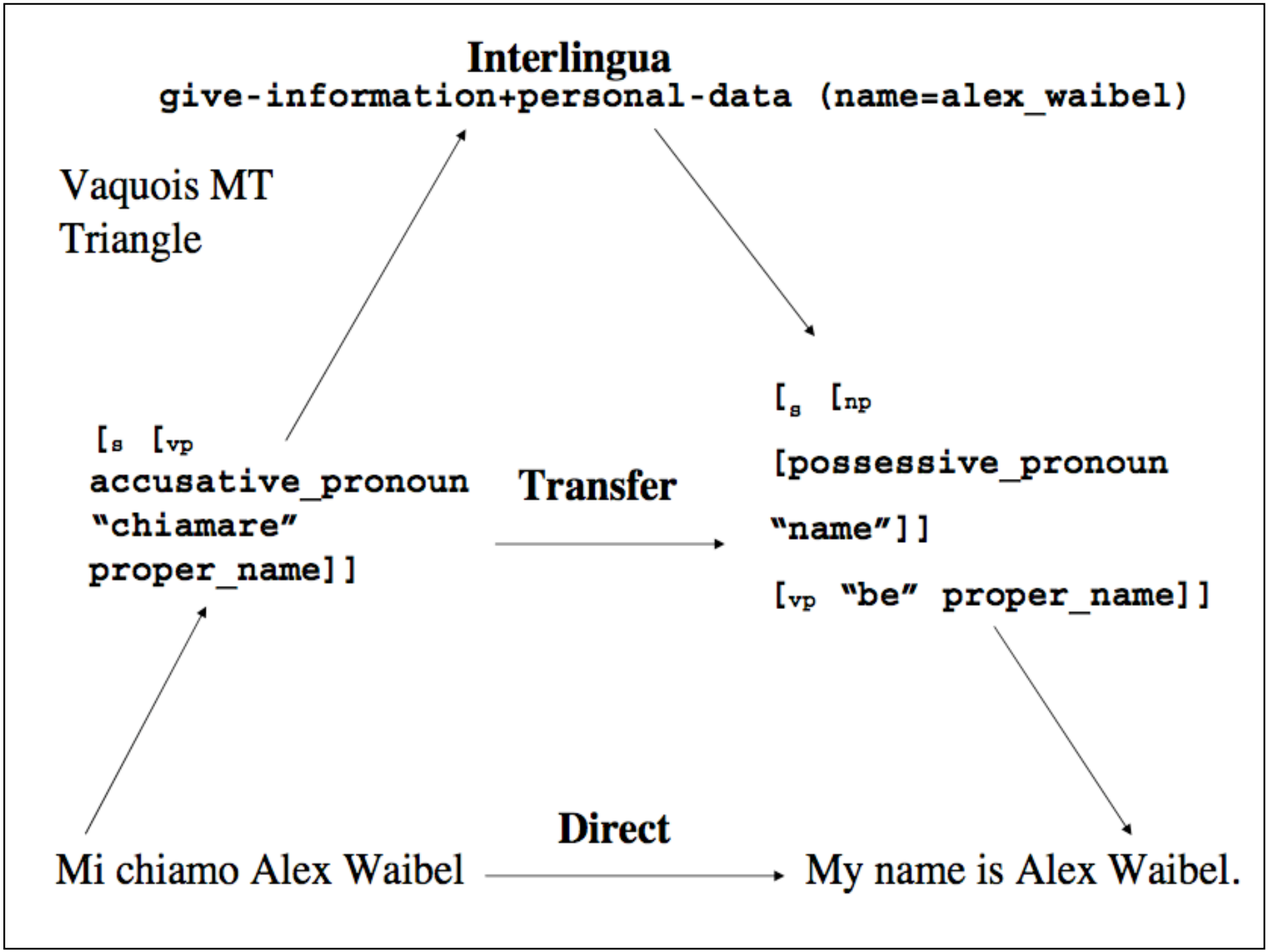
- They should work on a word-by-word basis and then reorganize the output to fit the target language. Typically for Romance to Germanic languages, the order of noun and adjective modifiers is different
- They are usually horizontally non-modular - no source vs. target language module -
- They are vertically modular: one module for each linguistic translation process, lexical, morphological, syntactic

Direct Systems

- They are usually domain limited
- Typical cases are Météo and Systran and many commercial systems like Weidner
- Systran translates 1/2 million pages each year for the European Commission (see the Europarl corpus made available for 27 languages)
- TAUM-Météo is translating public weather-forecasts from English to French in Canada from 1976 until today
- These are two successful examples of systems working with restriction on DOMAIN or/and type of language input which is CONTROLLED
- Danish PaTrans translating patent applications from English to Danish

Approaches to MT

- **Direct:**
  - Very little analysis of the source language.
- **Transfer:**
  - Analysis of the source language.
  - The structure of the source language input may not be the same as the structure of the target language sentence.
  - Transfer rules relate source language structures to target language structures.


Interlingua
give-information+personal-data (name=alex_waibel)
Vaquois MT
Triangle
[s [vp
accusative_pronoun
"chiamare"
proper_name]]
Transfer
[s [np
[possessive_pronoun
"name"]]
[vp "be" proper_name]]
Direct
Mi chiamo Alex Waibel
My name is Alex Waibel.

What is an interlingua?

- **Representation of meaning or speaker intention.**

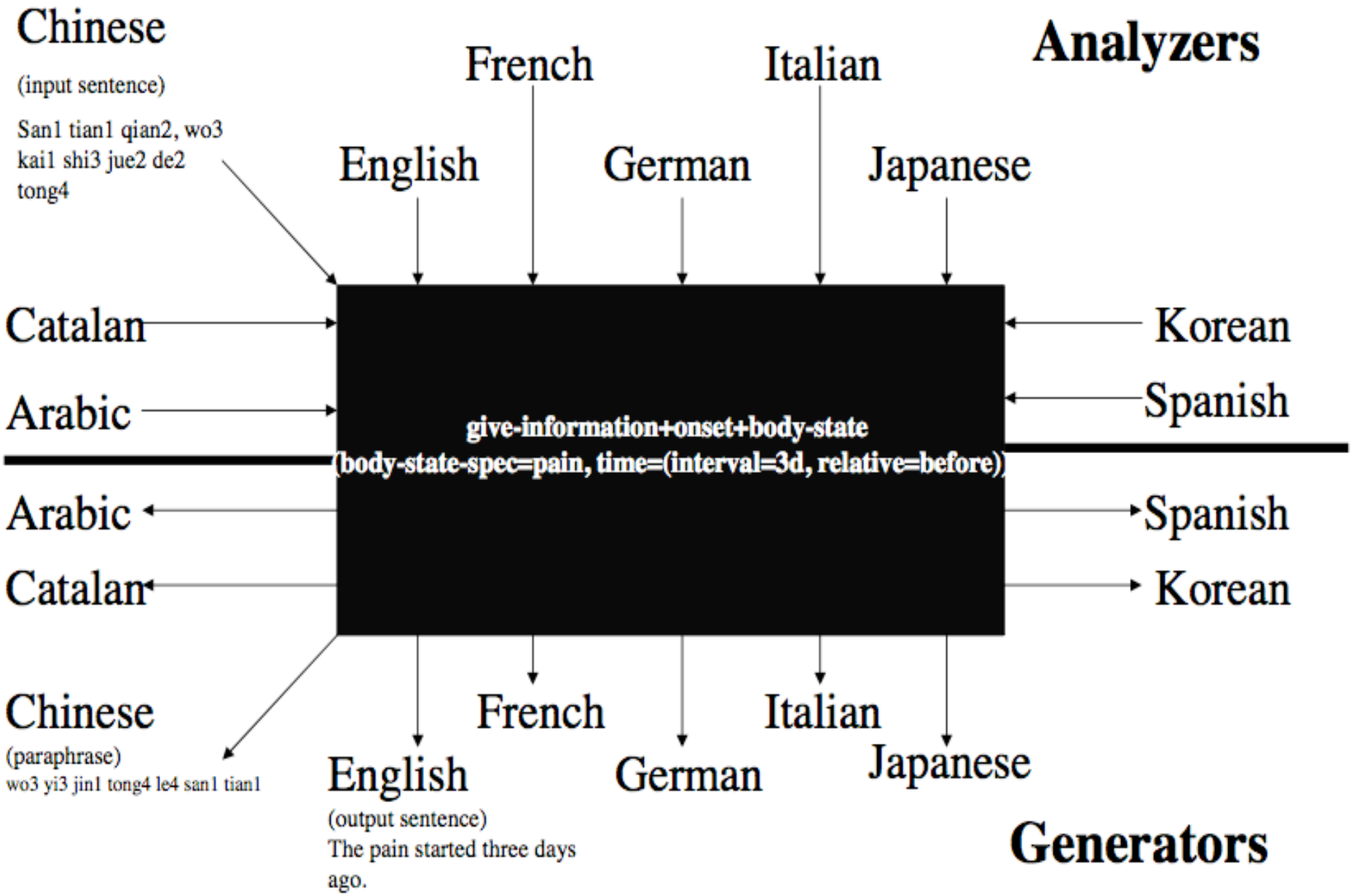
- **Sentences that are equivalent for the translation task have the same interlingua representation.**

The room costs 100 Euros per night.
The room is 100 Euros per night.
The price of the room is 100 Euros per night.

Note

- Some transfer systems may produce a more detailed meaning representation than some interlingua systems.
- The difference is whether translation equivalents in the source and target languages are related by a single canonical representation or not.

Multilingual Translation with an Interlingua



Multilingual translation with transfer

- Transfer-rules-1: Arabic-Catalan
- Transfer-rules-2: Catalan-Arabic
- Transfer-rules-3: Arabic-Chinese
- Transfer-rules-4: Chinese-Arabic
- Transfer-rules-5: Arabic-English
- Transfer-rules-6: English-Arabic
- Etc.

Advantages of Interlingua

- **Add a new language easily**
  - get all-ways translation to all previous languages by adding one grammar for analysis and one grammar for generation
- **Mono-lingual development teams**
- **Paraphrase**
  - Generate a new *source* language sentence from the interlingua so that the user can confirm the meaning

Disadvantages of Interlingua

- **"Meaning" is arbitrarily deep**
  - What level of detail do you stop at?
- **If it is too simple, meaning will be lost in translation.**
- **If it is too complex, analysis and generation will be too difficult.**
- **Should be applicable to all languages**
- **Human development time**

Interlingual MT Systems

- University of Maryland – Lexical Conceptual Structure (Dorr)
- Carnegie Mellon
  - Kantoo (Mitamura and Nyberg)
  - Nespole/C-STAR (Waibel, Levin, Lavie)
- UNL (Universal Networking Language)
- Microcosmos (Nirenburg)
- Verbmobil – Domain actions (Block)

Translation Mismatches

- Sentences that are translation-equivalents in two languages do not have the same syntactic structure or predicate-argument structure.
  - *I* like to swim
  - *Mi* piace nuotare

English He **likes** to swim.
German Er schwimmt **gern**.

English The baby **just** ate.
Spanish El bebé **acaba de** comer.

  - I swam *across* the river
  - Ho *attraversato* il fiume a nuoto

Lexicalization problems

- Germanic languages typically express manner (obligatory) as part of the verb and direction in the preposition of the prepositional phrase
- Romance languages express direction in the verb and manner in a (non-obligatory) participle or prepositional phrase

Lexicalization problems

English He swam across the channel.

Danish Han svømmede over kanalen.

Spanish Cruzó el canal nadando.

English He walked across the street.

Danish Han gik over gaden.

French Il traverse la rue a pied.

Italian Attraversa la strada a piedi.

Speech Acts: speaker intention vs literal meaning

"Can you pass the salt"

- **Literal meaning:** The speaker asks for information about the hearer's ability.
- **Speaker intention:** The speaker requests the hearer to perform an action.

Formulaic Utterances

- Good night.
- tisbaH cala xEr

  waking up on good

- Romanization of Arabic from CallHome Egypt

Language Neutrality

- **Comes from representing speaker intention rather than literal meaning for formulaic and task-oriented sentences.**

How about …
Why don't you…

suggestion

Could you tell me…
I was wondering…

request-information

Backside of Linguistic Knowledgeable Systems

- The language defined by the “grammars” at all levels of linguistic knowledge is never a complete grammar. Also it will never be exactly replicating the real language due to lack of perspicuity. Difficult to string together ***linguistic rules, lexical knowledge and domain related constraints.***
- In other cases ROBUSTNESS is the problem
- At a lexical level it means reducing the number of possible translation pairs

Transfer Systems

- In transfer as in direct translation, you need to have separate modules for each translation task
- Modules are thus more principled, since they represent comparative, linguistic knowledge
- This allows a distinction between procedural knowledge - related to such tasks as parsing, generation and transfer
- From declarative knowledge, which is typically the grammar of the language and the comparative knowledge.

Transfer Systems

- This means that you only need to have one parser, one generator and one transfer algorithm for the whole system
- Then you need one grammar for each language, both for analysis and generation
- And one comparative grammar for both directions of transfer.
- Grammars may be modularized into morphology, syntax, lexical specialized categories, collocates, multiwords, etc.

Transfer Systems

- In parsing you may want to have overcoverage, allowing the system to accept as many grammatical sentences as possible, plus some which are not and perhaps will only occur in specific contexts/domains/genres etc.
- On the contrary, generation needs to have a slight undercoverage, possibly preventing it from generating a few grammatical sentences, but making sure, that the sentences it does generate are grammatical.
- The same may apply to the transfer module.

Transfer Systems

- From a theoretical point of view there is one major advantage in the transfer-based approach. It is concerned primarily with questions of contrastive linguistic analysis, namely the comparison of the linguistic devices that the two languages employ to convey similar meanings.
- In the interlingua approach all the contrastive linguistics is hidden in the monolingual analysis and generation
- In the transfer system contrastive differences are captured in the transfer module: lexical substitution, structural change, features, etc.

Work Done on *STILVEN*

- Minimal Pairs
  - Due to Stress and Open/Closed e/o
  - Orthographical variants
- Corpus homogenization
- Dictionaries
- Documents collected/used
- English corpus fully parsed
- Base Lexicon (fully specified)

Work Done on *STILVEN*

- Frequency lists
- Multiwords - both Italian/English, English/Italian, Veneto/English
- Automatic MT based on MOSES
  - English/Veneto Parallel Aligned corpus
  - Translation Model for English/Veneto
- XLE grammar and parsing

Minimal Pairs and/or Orthographical Variants?

- Veneto as a spoken rathen than written language
  - Orthography is not well agreed upon
  - Word stress needs to marked explicitely
  - Degree of openness of vowels e-o must be marked explicitely
  - Users are led astray by the orthography of Italian (c/ch/q for K when the sound is /k/)

Minimal Pairs and/or Orthographical Variants?

- Orthographical variants
  - Four main recognized variants:
    - Central (Padovano spoken by most)
    - Western (Vicentino)
    - North-Eastern (mountain side)
    - Venetian
  - Due to sociological parameters like
    - Age, social role, center/periphery location
    - Geographical influence
    - Historical reasons

Corpus of Parallel Texts

- 69,000/67,000 tokens approximately
  - 20,000 are punctuation marks
- However, only 49,000 have been used
- 3 types or domains
  - Children stories and plays for children
  - History of the United States
  - Manuals and grammars

MOSES

- Open Source Toolkit for Statistical Machine Translation
- a factored phrase translation beam search decoder
- We propose to extend phrase-based statistical machine translation models using a factored representation.
- Current statistical MT approaches represent each word simply as their textual form.
- A factored translation approach replaces this representation with a feature vector for each word derived from a variety of information sources.
- These features may be the surface form, lemma, stem, part-of-speech tag, morphological information, syntactic, semantic or automatically derived categories, etc.

MOSES

- This representation is then used to construct statistical translation models that can be combined together to maximize translation quality. We expect that this representation will benefit MT in two ways:
- It allows better generalization over the data (e.g. treating "car" and "cars" unified in the translation model) - this is especially beneficial for morphological rich languages (such as Arabic, Finnish, Slavic languages), and for limited-data conditions.
- Syntactic representations can be used during reordering (which is often syntactically motivated), to ensure overall grammatical coherence (a weakness of current phrase-based models).

GIZA Ouput

Sentence pair (1846) source length 6 target length 6 alignment score : 6.46283e-09
te demo na man catarlo .
NULL ({ }) we'll ({ 1 2 }) help ({ 3 4 }) you ({ }) find ({ 5 }) it ({ }) . ({ 6 })

Sentence pair (1841) source length 7 target length 8 alignment score : 6.57471e-10
vuto dir ke xe on porta furtuna ?
NULL ({ 3 }) you ({ }) mean ({ 1 2 }) it's ({ 4 }) a ({ 5 }) lucky ({ 6 }) charm ({ 7 }) ? ({ 8 })

Sentence pair (72) source length 10 target length 12 alignment score : 7.02798e-13
a go idea ke xe ora de nar caxa , eh .
NULL ({ 1 4 }) i ({ 2 }) guess ({ 3 }) it's ({ 5 }) time ({ 6 }) to ({ 7 }) go ({ 8 }) home ({ 9 }) , ({ 10 }) huh ({ 11 }) . ({ 12 })

Dictionaries

- 9000 fully classified verbs of Veneto into English and Italian
- 38,000 translation pairs Veneto English with features and lexical category
- 3000 entries of Veneto unclassified

Lexicon Fully Specified

- It is organized into two separated files and it is referred to Italian
  - 1st file is the bilingual lexicon Veneto/Italian
  - 2nd file is the subcategorized lexicon of Italian
- It contains 1700 entries
  - Verbs, Nouns, Prepositions, Adverbs, Conjunctions, Adjectives, Pronouns

Frequency lexicon

- We ended up with 38,000 words
- 6,000 types
- All general texts had the following
  - 178,000 tokens
  - 18,000 types

FREQ ORDER

1294_el
1263_de
1253_£a
984_e
980_ke
957_a
747_xe
694_i
627_na
585_no
571_par
564_in
484_te
435_on

412_ga
405_da
404_se
349_so
339_£e
326_co
310_ma
289_me
270_sta
261_xera
228_ghe
195_la
189_del
188_go
178_mi
165_anca

Multiwords

```
mws(picked_up,tol_su).
mws(after_all,dopotuto).
mws(all_right,va_ben).
mws(and_so,e_anca).
mws(at_all,par_sogno).
mws(at_home,a_caxa).
mws(at_work,al_laoro).
mws(a_bit,na_scianta).
mws(a_bit,on_poco).
mws(a_few,on_par).
mws(a_little,na_scianta).
mws(a_lot,na_fraca).
mws(a_lot,na_gran).
mws(a_lot,on_mucio).
```

XLE or LFG compliant parsing

- PARC PARGRAM
- LFG formalism
- Fully specified lexicon
- Context-sensitive rules
  - Context-free with functional equations

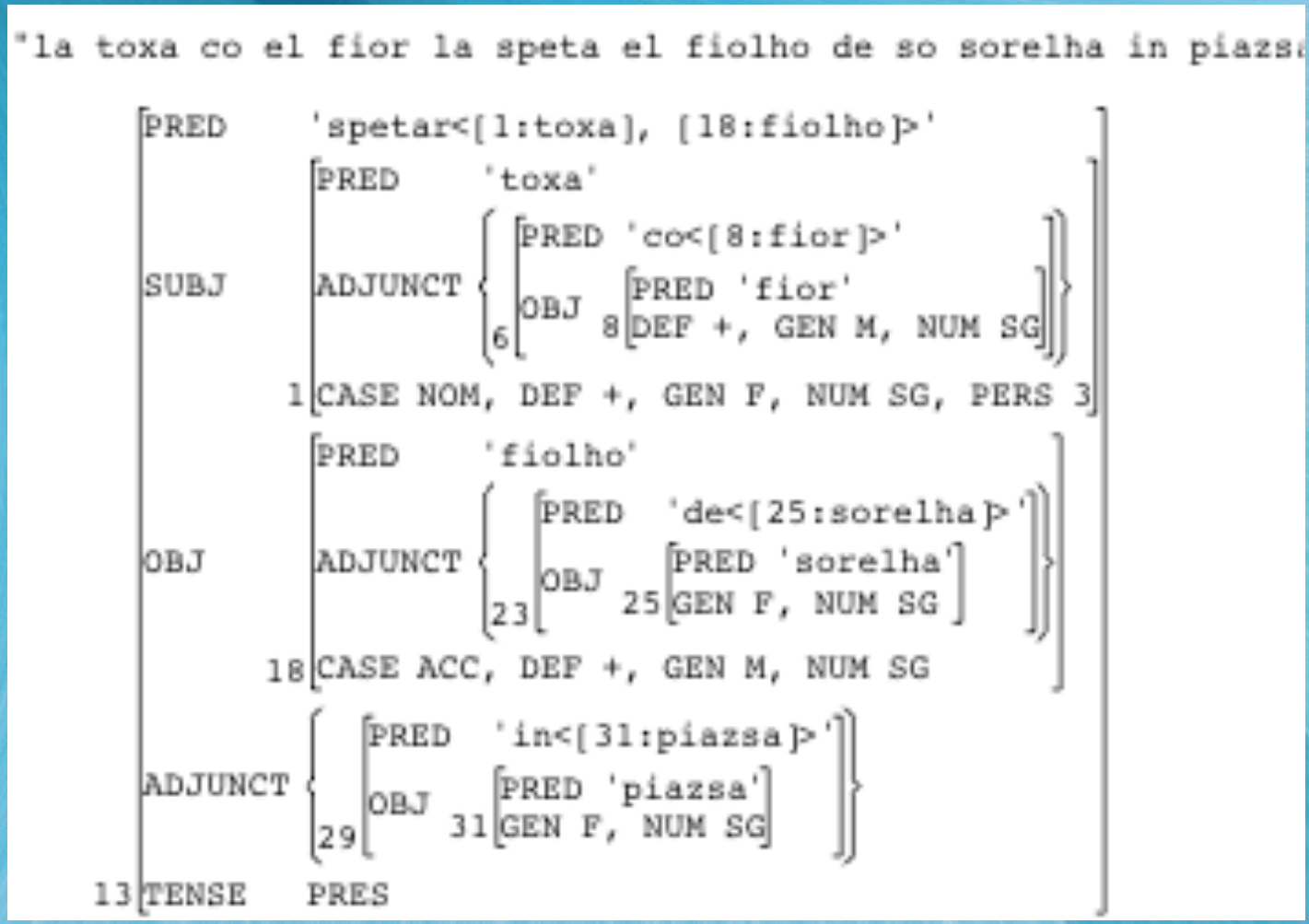
XLE or LFG compliant parsing

```
S --> {
          NP:   (^ SUBJ)=!
                (! CASE)=NOM;
          |
          e: (^ SUBJ PRED)='pro'
         }
         IP: ^=!.

 IP --> CLI-SUBJ
       (CLI-DAT)
         (AUX)
         VP.
```

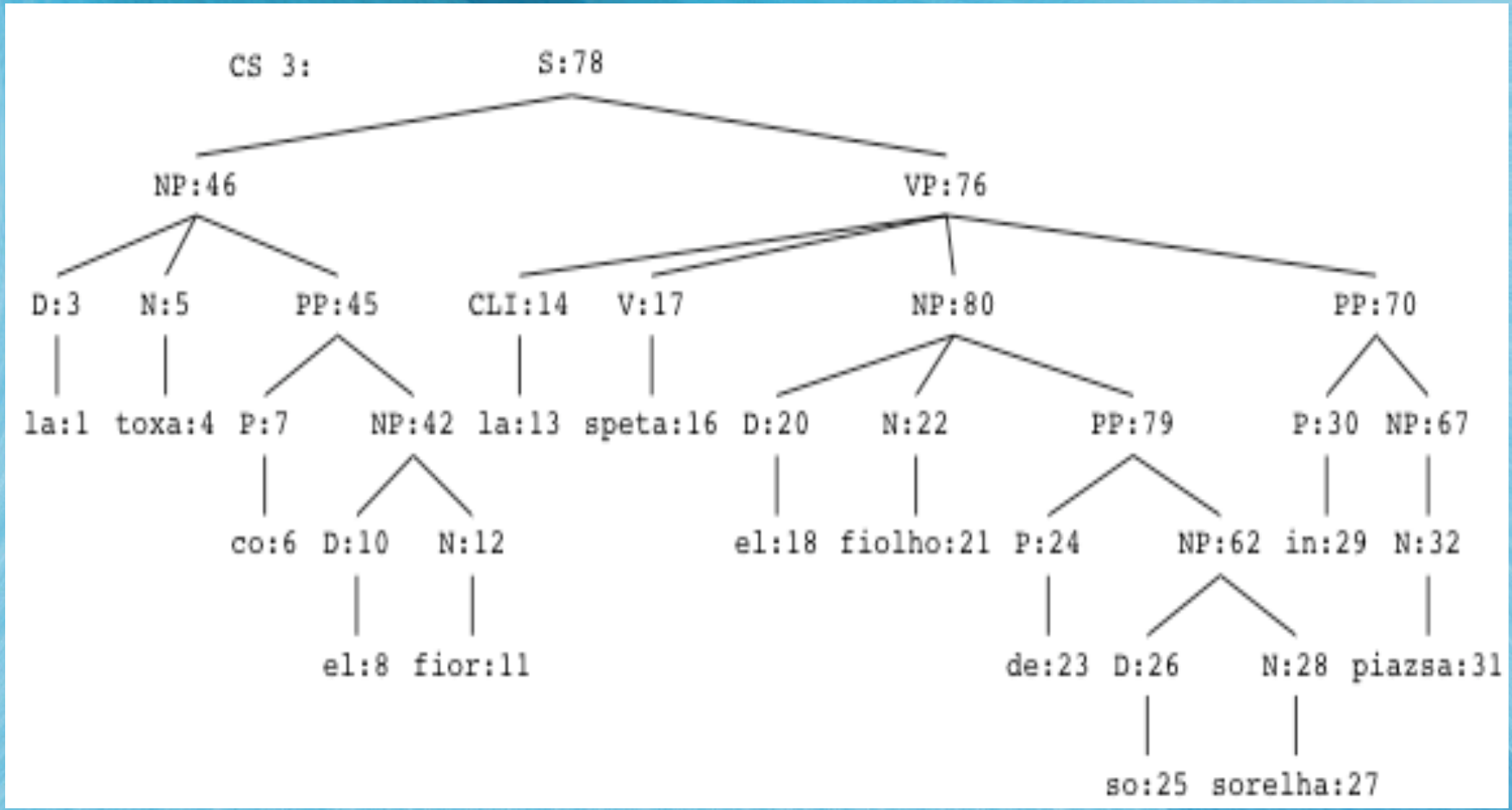

XLE or LFG compliant parsing

```
VP --> V: ^=!;
       {
         (NP:  (^ OBJ)=! (! CASE)=ACC;)
       |
       C S:    (^ COMP)=!
       }
         PP*: ! $ (^ADJUNCT).

  NP --> (D)
           N
       PP*: ! $ (^ADJUNCT).

  PP -->       P:      ^=!;
               NP:     (^ OBJ)=!.
```


XLE or LFG compliant parsing



XLE or LFG compliant parsing



TRANSFER MODULE

- We intend to organize a pipeline of modules that perform a complete transfer from Veneto to Italian and from Italian to English
- In this way we can exploit all information available and freely downloadable of parallel aligned texts for Italian and English

TRANSFER MODULE

- Translating Veneto to Italian and viceversa is feasible
- Structural differences are not many and can be easily spotted
- In this way, also orthographic variations and lexical variants can be treated

TRANSFER MODULE

- Veneto lacks past tense/passato remoto and uses perfect instead
- Wh- questions introduce a dummy BE and doubles the complementizer
- Clitic subjects may also be doubled
- Yes/no questions use clitic inversion

Some Examples

- Cosa facciamo adesso?
  - Cosa xé che faxemo deso?
- Glielo ho potuto dare
  - Go poduo dargheo
- Mi ha dovuto ricevere per forza
  - El me ga dovuo riséver par forsa
- I ragazzi si sono visti tutti in Tv
  - I tosi i se ga vardà tuti in Tv
- Giorgio mi deve parlare domattina
  - Giorgio el ga da parlarme doman matina

Conclusions

- We decided to go through an intermediate language, Italian
- This allows us to take advantage of quantitative data freely available
- Technically, we intend to use two approaches
  - The Transfer based approach
  - The AMT based approach using MOSES