# Minimal Dependency Translation:
# a Framework for CAT for Under-Resourced Languages

Michael Gasser
Indiana University, School of Informatics and Computing
Bloomington, Indiana, USA
gasser@indiana.edu

For under-resourced languages (URLs), the communities of speakers suffer from a lack of written material in their mother tongues. A partial solution to the problem is the translation of documents from other languages into the URLs. Computer-assisted translation (CAT) can speed up this process, but CAT systems require sizable translation memories, which are not available when one of the languages is under-resourced.

This paper describes an ongoing project to develop a lexical-grammatical framework for CAT with URLs as the target languages (TLs), relying on the grammatical resources and bilingual dictionaries that are available for many URLs. Called Minimal Dependency Translation (MDT), the framework is built on a lexicon of phrasal units called **groups**. Translation of a sentence results in an unordered set of translations of instantiated source-language (SL) groups.

Processing in MDT is illustrated below for the translation into Guarani of the Spanish sentence *no vamos a hablar con los maestros* ‘we aren’t going to speak with the teachers’ (1). The sentence is first subjected to POS tagging and morphological analysis, and a series of morphosyntactic transformation rules brings the input closer to TL structure (2). For example, the negator *no* and periphrastic future marker *vamos a* ‘we are going to’ are incorporated into the verb *hablar* ‘speak’, corresponding to Guarani morphology. Next the system searches for groups matching the input; three are shown (3). Two of these groups have heads that are lexemes rather than wordforms. For example, the group `<con $n>` matches sequences consisting of the preposition *con* ‘with’ followed by any noun. Next, constraint satisfaction is used to find a set of groups that covers the input sentence. In this process, group instantiations may be **merged**; in the example, the `$n` element in `<con $n>` unifies with the head of `<maestro_n>` ‘teacher’ to form a single dependency structure. Next TL groups are accessed for each selected SL group (4). Cross-linguistic feature agreement constraints in the group entries are applied (for example, TL verbs agree with SL verbs on the negation feature), and merged groups are merged for the TL (5). Thus, the `$n` element in `<$n ndive>` ‘with $n’, unifies with the head of `<mbo’ehára_n>` ‘teacher’. Finally, morphological generation is applied to the resulting TL lexemes and features (6). A single possible translation is shown for each SL phrase: `nañañe’ẽmo’ãi` ‘we will not speak’, `mbo’eharakuéra ndive` ‘with teachers’.

```
(1) No vamos a hablar con los maestros.
(2) hablar_v[t=fut,+neg,pn=1p]
    con maestro_n[+pl]
(3) <hablar_v>,<con $n>,<maestro_n>
(4) <ñe’ẽ_v>,<$n ndive>,<mbo’ehara_n>
(5) ñe’ẽ_v[t=fut,+neg,pn=1p],
    mbo’ehara_n[+pl] ndive
(6) nañañe'ẽmo'ãi; mbo'eharakuéra ndive
```

The goals of the project are (1) the development of a set of open-source tools for creating MDT implementations and (2) two functioning MDT implementations, one for Spanish–Guarani (http://guarani.soic.indiana.edu/mainumby/), the other for English–Amharic. The project began in 2016; following user testing in early 2018, the projected end date is late 2018. We are collaborating with the translation community in Paraguay through the Ateneo de Lengua y Cultura Guaraní and with the IT PhD Program at Addis Ababa University.

Ongoing research is concerned with methods for handling ambiguity (SL morphology and syntax, group assignment during constraint satisfaction, group translation) and for extending and correcting the lexicon-grammar based on user feedback and the limited bilingual corpora that are available.