# Developing Linguistically Informed Neural Net Machine Translation Systems

by
Yuan-Lu Chen

---

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF LINGUISTICS
In Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy
In the Graduate College
The University of Arizona

April 21, 2018

Get the official approval page
from the Graduate College
*before* your final defense.

## Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Signed: _____

**Chapter 1**

# What are glosses? Why are they golden representations of meanings?

## 1.1 Introduction: What are Glosses?

Interlinear Glossed Text is widely used in linguistic studies. The following is an example of Interlinear Glossed Text.

(1) Indonesian (Sneddon et al., 2012, p. 237)

Mereka di Jakarta sekarang. (*sentence of interest*)
they in Jakarta now (*gloss line: word-by-word gloss translation*)

'The are in Jakarta now.' (*English translation*)

A chunk of an interlinear glossed text has three lines. The first line is the sentence of interest. The second line is the gloss line, which is a word-by-word translation of the first line. And the third line a free English translation of the first line.

The conventional way to show the word-by-word translation from the first line to the gloss line is to use vertical alignment. In (1), '*Mereka*' is glossed as '*they*', '*di*' is glossed as '*in*', '*Jakarta*' is glossed as '*Jakarta*', and '*sekarang*' is glossed as '*now*'. These pairs are vertically aligned.

The gloss line also provides morphological information. Consider the following example:

(2) French

aux chevaux
to.ART.PL horse.PL

'to the horses'

The morphemes of a single word is linked by a '.'. The French word 'aux' is actually a combination of three separate morphemes[1]: '*to*', '*ART*', and '*PL*' and *'chevaux'* is decomposed into *'horse'* and *'PL'*. Bickel et al. (2008) compile a set of widely used conventions of IGT called the Leipzig Glossing Rules. Note that they are just guidelines of the formats of Interlinear Glossed Texts, so that Interlinear Glossed Texts can be more standardized.

The underlying intuition of Interlinear Glossed Text is that it provides an access to look into the subparts of a sentence. We may imagine the situation without the gloss line; then all we have is just the sentence and the English translation of that sentence. This will make it really hard to discuss the internal structure of the sentence. On the other hand, with the presence of the gloss line, with which each word is glossed and annotated, we then have a meta-representation in hand to discuss the grammatical properties of the sentence of interest.

An important note of the gloss line is that it is NOT raw linguistic data, and it is already processed. A linguist has already committed to some theory or some analysis on the sentence of interest when he or she transcribes the sentence into a gloss line, even if he or she tries to be as neutral as possible. As such, the question of what the gloss of a word is not trivial at all. Actually, sometimes a whole linguistic paper or thesis is to discuss and argue what the right gloss for a word is.

(3) Mandarin Chinese

Zhangsan **hen** gao
Zhangsan HEN tall

'Zhangsan is tall.'

For example, Grano (2008), Chen (2010), and Liu (2010) discuss the nature of the Mandarin Chinese word '*hen*' in the above example and what the right gloss should

[1] A morpheme is a smallest unit of meaning. For example, '*boys*' has two morphemes in it: '*boy*' and '*-s*', where '*-s*' is a plural marker. Sometimes, the morpheme boundary is not visible. For example '*went*' is composed of '*go*' and '*-ed*'.

be '*hen*'. In cases like this, how one glosses a word is not trivial at all, but determining what the gloss of word is requires a set of evidence and arguments.

## 1.2 The Golden Properties of Glosses

A system of meaning representations is decomposed of three components: a) meanings, b) representations, and c) a mapping between meanings and representations. The most ideal meaning representation system should be built with one-meaning-to-one-representation mappings; in other words, a meaning is mapped to one and only one representation. Natural languages fail to do so, given that synonyms and ambiguous words/phrases are ubiquitous in natural languages. On the other hand, gloss provides a mapping that is close to this ideal one-to-one mapping. Thus gloss should a better representation in term of representing meanings.

Theoretically, the claim that gloss representation is closer to the ideal one-to-one mapping than natural language representation is can be tested empirically. Let's imagine a set of special golden meta-linguistic semantic representations, which has the following property: each concept is mapped to one and one representation and each representation is mapped to one and one concept. With this imaginary golden semantic representation system, we may now compare Gaelic words and glosses. First, it is expected that each golden representation token will map to more natural language words than gloss items do.

(4) a. $golden_i \rightarrow \{Gaelic\_word_1, Gaelic\_word_2, \ldots\}_{golden_i}$

b. $golden_i \rightarrow \{gloss_1, gloss_2, \ldots\}_{golden_i}$

c. $|\{Gaelic\_word_1, Gaelic\_word_2, \ldots\}_{gold_i}| \geq |\{gloss_1, gloss_2, \ldots\}_{gold_i}|$

(4a) and (4b) represent a singe golden token may maps to multiple Gaelic words and glosses respectively. If we compare the size of them, it is expected that the set of Gaelic words is bigger than that of glosses, meaning that Gaelic words are more likely

to be homographs than glosses are. Section 1.2.1 will provide concrete examples to exemplify this property of glosses.

For the other direction, we may determine which one, Gaelic words or glosses, is more likely to be ambiguous.

(5) a. $Gaelic\_word_i \rightarrow \{golden_1, golden_2, \ldots\}_{Gaelic\_word_i}$

b. $gloss_i \rightarrow \{golden_1, golden_2, \ldots\}_{gloss_i}$

c. $|\{golden_1, golden_2, \ldots\}_{Gaelic\_word_i}| \geq |\{golden_1, golden_2, \ldots\}_{gloss_i}|$

(5a) and (5b) show the mappings from a Gaelic word to different concepts and the mappings from a gloss to different concepts respectively. (5c) is the expectation that Gaelic words are more likely to be ambiguous than glosses are. Section 1.2.2 will report concrete examples to show that glosses are less likely to be ambiguous.

To run statistical experiments to confirm the truth of (4c) and (5c) is the way to empirically support the claim that glosses closer to the golden representations than Gaelic words are. However, in reality, this is an impossible experiment to conduct, because there are no such golden representation[2]. In spite of the impossibility of conducting statistical experiments, we may still use some examples to show the intuition that glosses are better representations than natural languages are. The following sections describes how glosses cluster words with different forms but with the same meaning, and how glosses represent words with same form but with different meanings with different representations.

### 1.2.1 Glosses Cluster Different Words with the Same Meanings (Synonyms) Into a Single representation

Gloss collapses words with different forms with the same meanings into a single gloss. In natural languages, the morphology of a word (i.e. the form of a word) may be

[2]It would solve the puzzle of semantics if one should be able to build the set of special golden meta-linguistic semantic representations, and the mappings between the golden representations to natural languages.

sensitive to the phonological environments and changing into different forms. Consider the following the indefinite article in the English examples:

(6) John ate **an** apple.
John eat.past **INDF_ART** apple

(7) John ate **a** banana.
John eat.past **INDF_ART** banana

In the above example, *an* and *a* have the identical meaning[3]. In English, the same concept is realized as two representations, *a* or *an*, while in the gloss representation the one concept is neatly represented as *INDF_ART* (indefinite article).

Critically, synonyms like the English *a* and *an* commonly occur in many other natural languages if not in all languages. The definite article in the language of interest, Scottish Gaelic, is another example to show the noisiness of natural language representations. Consider the definite article in the following Gaelic examples.

(8) tha mi a' sireadh **an** leabhair bhig ghuirm
be-PRES-IND 1S PROG searching-VN **ART** book-G small-G blue-G
'I am looking for the small blue book' (Lamb, 2001, p. 29)

(9) **am** fear mòr
**ART** man big
'a big man' (Lamb, 2001, p. 31)

(10) thuit **a'** chlach air cas mo mhnà
fall-PAST **ART** stone on foot 1S-POSS wife-G
'the stone fell on my wife's foot' (Lamb, 2001, p. 30)

(11) doras **na** sgoile(adh)
door-N **ART** school-G
'the door of the school' (Lamb, 2001, p. 29)

[3]Semantically, *an* and *a* are existential quantifiers, which declare that a member of a set exists in the world. In formal semantics, *an* and *a* may be defined as follows: $\exists\lambda P[P(x)]$. In the current example, *apple* and *banana* will instantiate $P$ in the formula, and the meanings will be 'an apple exists' and 'a banana exists'. Kratzer and Heim (1998) would be a nice introduction for interested readers to see how linguists, specifically semanticians, define, decompose, and compose meanings of languages formally.

(12) a chuir air dòigh **nan** àiridhean a-muigh a rubh' Eubhal agus an oidhche seo
to put-INF on order **ART** sheilings out-LOC to point Eaval and ART night this
'the girls big house' (Lamb, 2001, p. 100)

(13) fèis **nam** bàrd
festival **ART** poet.PL.GEN
'festival of the poets' (Lamb, 2001, p. 107)

The definite article in Scottish Gaelic may be realized as the following forms: as *an*, *am*, *a'*, *na*, *nan* or *nam*. The alternation is determined by the case, gender and number of noun phrase that it modifies, and additionally the phonological property of the word following it also changes the form of the definite article (Lamb, 2001). All these different realizations refer to the same concept, the definite article. Again, the gloss notation nicely clusters them together as *ART*.

In Mandarin Chinese, similar patterns are found. Consider classifiers in the following examples:

(14) Yani mai-le **{pi/*tou}** ma , Lulu mai-le **{*pi/tou}** zhu.
Yani buy-PRF CL/CL horse , Lulu buy-PRF CL/CL pig
'Yani bought a horse and Lulu bought a pig.' (Zhang, 2013, p. 136)

In Zhang (2013), the classifier like *pi* and *tou* is a type of *indivual classifier* which co-occurs with countable nouns, like *ma*, 'horse', and *zhu*, 'pig', and this type of classifier is the head of *UNIT Phrase*. *Pi* and *tou* actually have the same semantics and the syntactic function; however, they are realized in different forms, specifically the form of which has to agree with the noun following it (i.e. *pi* goes with *ma*; *tou* goes with *zhu*). Here the gloss, *CL*, unifies the two forms of the same meaning.

Gloss collapses synonyms in natural languages. Learning the general distribution of the article and all its different forms is a challenge for the MT system, but the glossing information should make this easier.

### 1.2.2 Glosses Distinguish Homographs' Different Meanings

In natural languages, there are cases when a single form denotes distinct concepts. Words with this property are termed as homographs. Consider the word *for* in following English examples:

(15) a. I intended **for** Jenny to be present.

b. **For** Jenny, I intended to be present. (Adger, 2003, p.306-307)

*For* in (15a) and (15b) has the same form but different meanings. Specifically, *for* in (15a) is a complementizer with its part of speech being *C*, and it heads the non-finite clause *Jenn to be present*; on the other hand *for* (15b) is a preposition, which takes a Determiner Phrase, *Jenny*, as its benefactive argument.

The Scottish Gaelic word *a'* in the following examples also has different meanings.

(16) tha mi **a'** sireadh an leabhair bhig ghuirm.
be-PRES-IND 1S **PROG** searching-VN ART book-G small-G blue-G
'I am looking for the small blue book.' (Lamb, 2001, p. 29)

(17) thuit **a'** chlach air cas mo mhnà.
fall-PAST **ART** stone on foot 1S-POSS wife-G
'the stone fell on my wife's foot.' (Lamb, 2001, p. 30)

Critically *a'* in (16) is a progressive aspect marker while in (17) the some form denotes to definite article. Again, the semantic difference is preserved in the gloss representations but not in natural language words. The gloss data also provides hierarchical (non-linear) syntactic parsing information.

### 1.2.3 Glosses are Sensitive to Hierarchical Structures in Natural Language Sentences

Before I introduce how gloss information is linked to hierarchical structures, it is necessary to emphasize the importance of hierarchical structures in natural languages. In this section, I will first review some linguistic arguments for why and how semantics

and syntax of languages[4] are all about hierarchical structures instead of linear word orders. Then I will link gloss to hierarchical structures.

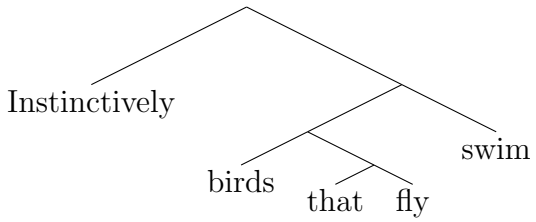
It is well-argued in linguistics that the syntax and semantics of natural languages are determined by hierarchical structures instead of linear orders of words, and essentially it is the sensitivity of hierarchical structures that distinguishes human natural languages from other animal communications (Berwick and Chomsky, 2015).

Semantics is determined by hierarchical structures instead of linear orders. Berwick and Chomsky (2015, p. 117) use the following simple example to demonstrate this property of natural languages:

(18) Instinctively birds that fly swim.

In the example above, *instinctively* is linearly close to *fly* than *swim*; however, it unambiguously modifies *swim* instead of *fly*. The reason for this is the hierarchical structures (Berwick and Chomsky, 2015, p. 117):

(19)




In (19) it is shown that *fly* is more embedded than *swim*, and thus it is hierarchically further away from *instinctively*. So, *instinctively* can only modify *swim* instead of *fly*.

Syntax is also all about hierarchical structures. Consider the following sentence:

(20) a. Birds that $\text{can}_1$ fly $\text{can}_2$ swim.

b. *$\text{Can}_1$ birds that fly $\text{can}_2$ swim?

c. $\text{Can}_2$ birds that $\text{can}_1$ fly swim?

[4]When it turns to the sound aspect of languages, Phonetics is more about linear order, but Phonology is still sensitive to hierarchical structures just like syntax and semantics.

(20a) is a declarative sentence. To derive an interrogative sentence from it, the auxiliary needs to be moved; however only $can_2$ can be moved but not $can_1$ even $can_1$ is linearly close to the sentence initial position. Again, it is all because of the hierarchical structures. $Can_2$ is in the matrix clause while $can_1$ is in the embedded relative clause.

Glosses, on the other hand, are sensitive to the internal hierarchical structures or constituency of sentences. They provide more clues of the internal hierarchical structures or constituency of sentences than natural language words. Consider the following examples, modified from (15):

(21) For as *complementizer* (glossed as *complementizer*)

a. I intended **for** [Jenny] to be present.

b. I intended **for** [the girl] to be present.

c. I intended **for** [the little girl] to be present.

d. I intended **for** [the little girl who wants to eat some ice scream] to be present.

(22) For as *preposition* (glossed as *preposition*)

a. **For** [Jenny], I intended to be present.

b. **For** [the girl], I intended to be present.

c. **For** [the little girl], I intended to be present.

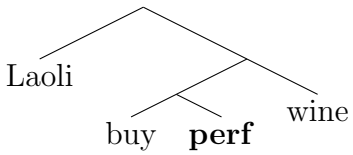
d. **For** [the little girl who wants to eat some ice scream], I intended to be present.

Linear length of the argument of *for* (i.e. the sequences in the square brackets) does not have any effect in determining what the gloss is, and instead it is the hierarchical structures that determines what the gloss is. Then the form of gloss hints to the internal structures of the sentence.

A even more dramatic example comes from Mandarin Chinese. A single sequences of words may have distinctive meanings because of different parses, and the difference of parses is marked by the differences of glosses. In the following examples, the sentence '*Lao3Li3 mai3 hao3 jiu3*'[5] may have two distinct meanings depending on the status of '*hao3*'.

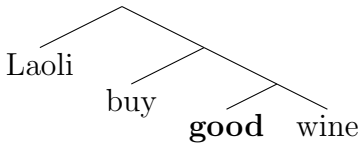
(23) a. Lao3Li3 mai3 hao3 jiu3
Laoli buy **Perf** wine
'Laoli bought a wine'

b.




(24) a. Lao3Li3 mai3 hao3 jiu3
Laoli buy **good** wine
'Laoli buys a good wine'

b.




In sentence (23), '*hao3*' goes with the verb '*mai3*'; as such '*hao3*' is interpreted as a Perfective marker and glossed as '*Perf*'; on the other hand, in (24), '*hao3*' goes with the noun '*jiu3*' and works as an adjective modifying '*jiu3*', and it is glossed as '*good*'.

With all the examples above, we have showed that gloss lines provide more clues of the the internal structures of the sentences are than natural language words do.

[5]These specific examples are extensively discussed in Mandarin Chinese Tone Sandhi literature (e.g. Cheng (1973); Mei (1991); Shih (1997); Wang and Lin (2011)). Critically, the constituency plays a role in Mandarin Chinese Tone Sandhi.

## 1.3 Conclusion: What Is a Gloss Line and Why Do They Matter?

The gloss line is like a linguistic version of 'word embedding'. A natural language word is first converted to a gloss, which is readable for linguists. Also we may view a gloss line as an artificial sentence using the purified 'gloss words', a meaning representation with which one meaning is mapped to one and only one representation. It is a useful and widely used annotation algorithm that requires linguistic knowledge. Given the properties of gloss data, it can be a very useful data for machine translation. Moreover, gloss data is widely used in linguistics literature, so data is already out there and all we need to do to clean the data. A loose end here is that, even if all the arguments should be sound, we still have no statistical evidence to show the usefulness of the gloss data. Chapter **??** and **??** close this loose end, in which I will report machine translation experiments using gloss data.

# Bibliography

Adger, David (2003), *Core syntax: A minimalist approach*, volume 33. Oxford University Press Oxford.

Berwick, Robert C and Noam Chomsky (2015), *Why only us: Language and evolution*. MIT press.

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath (2008), "The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses." *Revised version of February*.

Chen, Yuan-Lu (2010), *Degree Modification and Time Anchoring in Mandarin*. Ph.D. thesis.

Cheng, Chin-Chuan (1973), *A synchronic phonology of Mandarin Chinese*, volume 4. Walter de Gruyter.

Chomsky, Noam (2006), *Language and mind*. Cambridge University Press.

Grano, Thomas (2008), "Mandarin hen and the syntax of declarative clause typing." *Unpublished manuscript. Accessed online:< http://home. uchicago. edu/˜ tgrano/grano_hen. pdf>. First accessed*, 4.

Kratzer, Angelika and Irene Heim (1998), *Semantics in generative grammar*. Blackwell Oxford.

Lamb, William (2001), *Scottish Gaelic*, volume 401. Lincom Europa.

Liu, Chen-Sheng Luther (2010), "The positive morpheme in chinese and the adjectival structure." *Lingua*, 120, 1010–1056.

Mei, Tsu-Lin (1991), "Tone sandhi and morphological relics." *Journal of Chinese Linguistics Monograph Series*, 454–471.

Shih, Chilin (1997), "Mandarin third tone sandhi and prosodic structure." *Linguistic Models*, 20, 81–124.

Sneddon, James Neil, K Alexander Adelaar, Dwi N Djenar, and Michael Ewing (2012), *Indonesian: A comprehensive grammar*. Routledge.

Wang, Chiung-Yao and Yen-Hwei Lin (2011), "Variation in tone 3 sandhi: The case of prepositions and pronouns." In *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, 138–155.

Zhang, Niina Ning (2013), *Classifier Structures in Mandarin Chinese*, volume 263. Walter de Gruyter.