Meaning Unit Segmentation in English and Chinese: a New Approach to Discourse Phenomena

Jennifer Williams [†1,2], **Rafael Banchs**[2], **and Haizhou Li**[2]

[1]Department of Linguistics, Georgetown University, Washington, D.C., USA
[2]Institute for Infocomm Research, 1 Fusionpolis Way, Singapore
jaw97@georgetown.edu {rembanchs,hli}@i2r.a-star.edu.sg

Abstract

We present a new approach to dialogue processing in terms of "meaning units". In our annotation task, we asked speakers of English and Chinese to mark boundaries where they could construct the maximal concept using minimal words. We compared English data across genres (news, literature, and policy). We analyzed the agreement for annotators using a state-of-the-art segmentation similarity algorithm and compared annotations with a random baseline. We found that annotators are able to identify meaning units systematically, even though they may disagree on the quantity and position of units. Our analysis includes an examination of phrase structure for annotated units using constituency parses.

1 Introduction

When humans translate and interpret speech in real-time, they naturally segment speech in "minimal sense units" (Oléron & Nanpon, 1965; Benítez & Bajo, 1998) in order to convey the same information from one language to another as though there were a 1-to-1 mapping of concepts between both languages. Further, it is known that people can hold up to 7+/- 2 "chunks" of information in memory at a time by creating and applying meaningful organization schemes to input (Miller, 1956). However, there is no definitive linguistic description for the kind of "meaning units" that human translators create (Signorelli et al., 2011; Hamon et al., 2009; Mima et al., 1998).

The ability to chunk text according to units of meaning is key to developing more sophisticated machine translation (MT) systems that operate in real-time, as well as informing discourse processing and natural language understanding (NLU) (Kolář, 2008). We present an approach to discourse phenomena to address Keller's (2010) call to find a way to incorporate "cognitive plausibility" into natural language processing (NLP) systems. As it has been observed that human translators and interpreters naturally identify a certain kind of "meaning unit" when translating speech in real-time (Oléron & Nanpon, 1965; Benítez & Bajo, 1998), we want to uncover the features of those units in order to automatically identify them in discourse.

This paper presents an experimental approach to annotating meaning units using human annotators from Mechanical Turk. Our goal was to use the results of human judgments to inform us if there are salient features of meaning units in English and Chinese text. We predicted that human-annotated meaning units should systematically correspond to some other linguistic features or combinations of those features (i.e. syntax, phrase boundaries, segments between stop words, etc.). We are interested in the following research questions:

- At what level of granularity do English and Chinese speakers construct meaning units in text?
- Do English and Chinese speakers organize meaning units systematically such that meaning unit segmentations are not random?
- How well do English and Chinese speakers agree on meaning unit boundaries?
- Are there salient syntactic features of meaning units in English and Chinese?
- Can we automatically identify a 1-to-1 mapping of concepts for parallel text, even if there is paraphrasing in one or both languages?

[†] Now affiliated with Massachusetts Institute of Technology Lincoln Laboratory.

Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), pages 1–9,
Sofia, Bulgaria, August 9, 2013. ©2013 Association for Computational Linguistics

While we have not built a chunker or classifier for meaning unit detection, it is our aim that this work will inform how to parse language systematically in a way that is human-understandable. It remains to be seen that automatic tools can be developed to detect meaning units in discourse. Still, we must be informed as to what kinds of chunks are appropriate for humans to allow them to understand information transmitted during translation (Kolář, 2008). Knowledge about meaning units could be important for real-time speech processing, where it is not always obvious where an utterance begins and ends, due to any combination of natural pauses, disfluencies and fillers such as "like, um..". We believe this work is a step towards creating ultra-fast human-understandable simultaneous translation systems that can be used for conversations in different languages.

This paper is organized as follows: Section 2 discusses related work, Section 3 describes the segmentation similarity metric that we used for measuring annotator agreement, Section 4 describes our experiment design, Section 5 shows experiment results, Section 6 provides analysis, and Section 7 discusses future work.

2 Related Work

At the current state of the art, automatic simultaneous interpretation systems for speech function too slowly to allow people to conduct normal-paced conversations in different languages. This problem is compounded by the difficulty of identifying meaningful endpoints of utterances before transmitting a translation. For example, there is a perceived lag time for speakers when trying to book flights or order products over the phone. This lag time diminishes conversation quality since it takes too long for each speaker to receive a translation at either end of the system (Paulik et al., 2009). If we can develop a method to automatically identify segments of meaning as they are spoken, then we could significantly reduce the perceived lag time in real-time speech-to-speech translation systems and improve conversation quality (Baobao et al., 2002; Hamon et al., 2009).

The problem of absence of correspondence arises when there is a lexical unit (single words or groups of words) that occurs in L1 but not in L2 (Lambert et al., 2005). It happens when words belonging to a concept do not correspond to phrases that can be aligned in both languages. This problem is most seen when translating speech-to-speech in real-time. One way to solve this problem is to identify units for translation that correspond to concepts. A kind of meaning unit had been previously proposed as *information units* (IU), which would need to be richer than semantic roles and also be able to adjust when a mistake or assumption is realized (Mima et al., 1998). These units could be used to reduce the explosion of unresolved structural ambiguity which happens when ambiguity is inherited by a higher level syntactic structure, similar to the use of constituent boundaries for transfer-driven machine translation (TDMT) (Furuse et al., 1996).

The human ability to construct concepts involves both bottom-up and top-down strategies in the brain. These two kinds of processes interact and form the basis of comprehension (Kintsch, 2005). The construction-integration model (CI-2) describes how meaning is constructed from both long-term memory and short-term memory. One of the challenges of modeling meaning is that it requires a kind of *world-knowledge* or *situational knowledge*, in addition to knowing the meanings of individual words and knowing how words can be combined. Meaning is therefore constructed from long-term memory – as can be modeled by latent semantic analysis (LSA) – but also from short-term memory which people use *in the moment* (Kintsch & Mangalath, 2011). In our work, we are asking annotators to construct meaning from well-formed text and annotate where units of meaning begin and end.

3 Similarity Agreement

We implemented *segmentation similarity* (S) from Fournier and Inkpen (2012). Segmentation similarity was formulated to address some gaps of the *WindowDiff* (WD) metric, including unequal penalty for errors as well as the need to add padding to the ends of each segmentation (Pevzner & Hearst, 2002). There are 3 types of segmentation errors for (S), listed below:

1. s_1 contains a boundary that is off by n potential boundaries in s_2
2. s_1 contains a boundary that s_2 does not, or
3. s_2 contains a boundary that s_1 does not

These three types of errors are understood as *transpositions* in the case of error type 1, and as *substitutions* in the case of error types 2 and 3. Note that there is no distinction between insertions and deletions because neither of the segmentations are considered reference or hypothesis. We show the specification of (S) in (1):

$$S_{(si1,si2)} = \frac{\mathbf{t} \cdot mass(i) - \mathbf{t} - d_{(si1,si2,T)}}{\mathbf{t} \cdot mass(i) - \mathbf{t}} \quad (1)$$

such that S scales the cardinality of the set of boundary types **t** because the edit distance function $d_{(si1,si2,T)}$ will return a value for potential boundaries of $[0, \mathbf{t} \cdot mass(i)]$ normalized by the number of potential boundaries per boundary type. The value of $mass(i)$ depends on task, in our work we treat mass units as number of words, for English, and number of characters for Chinese. Since our annotators were marking only units of meaning, there was only one boundary type, and $(\mathbf{t} = 1)$. The distance function $d_{(si1,si2,T)}$ is the edit distance between segments calculated as the number of boundaries involved in transposition operations subtracted from the number of substitution operations that could occur. A score of 1.0 indicates full agreement whereas a score of 0 indicates no agreement.

In their analysis and comparison of this new metric, Fournier and Inkpen (2012) demonstrated the advantages of using (S) over using (WD) for different kinds of segmentation cases such as maximal/minimal segmentation, full misses, near misses, and segmentation mass scale effects. They found that in each of these cases (S) was more stable than (WD) over a range of segment sizes. That is, when considering different kinds of misses (false-positive, false-negative, and both), the metric (S) is less variable to internal segment size. These are all indications that (S) is a more reliable metric than (WD).

Further, (S) properly takes into account chance agreement - called *coder bias* - which arises in segmentation tasks when human annotators either decide not to place a boundary at all, or are unsure if a boundary should be placed. Fournier and Inkpen (2012) showed that metrics that follow (S) specification reflect most accurately on coder bias, when compared to mean pairwise $1 - WD$ metrics. Therefore we have decided to use segmentation similarity as a metric for annotator agreement.

4 Experiment Design

This section describes how we administered our experiment as an annotation task. We surveyed participants using Mechanical Turk and presented participants with either English or Chinese text. While the ultimate goal of this research direction is to obtain meaning unit annotations for speech, or transcribed speech, we have used well-structured text in our experiment in order to find out more about the potential features of meaning units in the simplest case.

4.1 Sample Text Preparation

Genre: Our text data was selected from three different genres for English (news, literature, and policy) and one genre for Chinese (policy). We used 10 articles from the Universal Declaration of Human Rights (UDHR) in parallel for English and Chinese. The English news data (NEWS) consisted of 10 paragraphs that were selected online from www.cnn.com and reflected current events from within the United States. The English literature data (LIT) consisted of 10 paragraphs from the novel Tom Sawyer by Mark Twain. The English and Chinese UDHR data consisted of 12 parallel paragraphs from the Universal Declaration of Human Rights. The number of words and number of sentences by language and genre is presented below in Table 1.

Preprocessing: To prepare the text samples for annotation, we did some preprocessing. We removed periods and commas in both languages, since these markings can give structure and meaning to the text which could influence annotator decisions about meaning unit boundaries. For the English data, we did not fold to lowercase and we acknowledge that this was a design oversight. The Chinese text was automatically segmented into words before the task using ICTCLAS (Zhang et al., 2003). This was done in order to encourage Chinese speakers to look beyond the character-level and word-level, since word segmentation is a well-known NLP task for the Chinese language. The Chinese UDHR data consisted of 856 characters. We placed checkboxes between each word in the text.

4.2 Mechanical Turk Annotation

We employed annotators using Amazon Mechanical Turk Human Intelligence Tasks (HITs). All instructions for the task were presented in En-

Language and Genre	# words	# Sentences
Chinese UDHR	485	20
English NEWS	580	20
English LIT	542	27
English UDHR	586	20

Table 1: Number of words and sentences by language and genre.

glish. Each participant was presented with a set of numbered paragraphs with a check-box between each word where a boundary could possibly exist. In the instructions, participants were asked to check the boxes between words corresponding to the boundaries of meaning units. They were instructed to create units of meaning larger than words but that are also the "maximal concept that you can construct that has the minimal set of words that can be related to each individual concept"[1]. We did not provide marked examples to the annotators so as to avoid influencing their annotation decisions.

Participants were given a maximum of 40 minutes to complete the survey and were paid USD $1.00 for their participation. As per Amazon Mechanical Turk policy, each of the participants were at least 18 years of age. The annotation task was restricted to one task per participant, in other words if a participant completed the English NEWS annotation task then they could not participate in the Chinese UDHR task, etc. We did not test any of the annotators for language aptitude or ability, and we did not survey language background. It is possible that for some annotators, English and Chinese were not a native language.

5 Results

We omitted survey responses for which participants marked less than 30 boundaries total, as well as participants who completed the task in less than 5 minutes. We did this in an effort to eliminate annotator responses that might have involved random marking of the checkboxes, as well as those who marked only one or two checkboxes. We decided it would be implausible that less than 30 boundaries could be constructed, or that the task could be completed in less than 5 minutes, considering that there were several paragraphs and sentences for each dataset. After we removed those responses, we had solicited 47 participants for English NEWS, 40 participants for English LIT, 59 participants for English UDHR, and 10 participants for Chinese UDHR. The authors acknowledge that the limited sample size for Chinese UDHR data does not allow a direct comparison across the two languages, however we have included it in results and analysis as supplemental findings and encourage future work on this task across multiple languages. We are unsure as to why there was a low number of Chinese annotators in this task, except perhaps the task was not as accessible to native Chinese speakers because the task instructions were presented in English.

5.1 Distributions by Genre

We show distributions of number of annotators and number of units identified for each language and genre in Figures 1 – 4. For each of the language/genres, we removed one annotator because the number of units that they found was greater than 250, which we considered to be an outlier in our data. We used the Shapiro-Wilk Test for normality to determine which, if any, of these distributions were normally distributed. We failed to reject the null hypothesis for Chinese UDHR ($p = 0.373$) and English NEWS ($p = 0.118$), and we rejected the null hypothesis for English LIT ($p = 1.8X10^{-04}$) and English UDHR ($p = 1.39X10^{-05}$).

Dataset	N	Avg Units	Avg Words/Unit
Chinese UDHR	9	70.1	–
English NEWS	46	84.9	6.8
English LIT	39	85.4	6.3
English LIT G1	26	66.9	8.1
English LIT G2	13	129.0	4.2
English UDHR	58	90.1	6.5
English UDHR G1	17	52.2	11.2
English UDHR G2	19	77.3	7.6
English UDHR G3	22	132.2	4.4

Table 2: Number of annotators (N), average number of units identified, average number of words per unit identified, by language and genre.

Since the number of units were not normally distributed for English LIT and English UDHR,

[1] The definition of "meaning units" we provide is very ambiguous and can justify for different people understanding the task differently. However, this is part of what we wanted to measure, as giving a more precise and operational definition would bias people to some specific segmentation criteria.

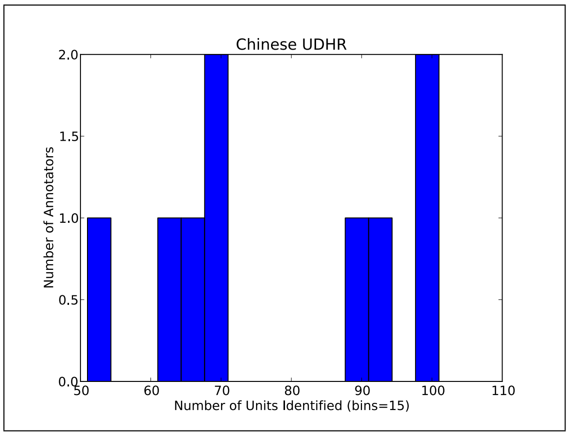

Figure 1: Distribution of total number of annotations per annotator for Chinese UDHR.

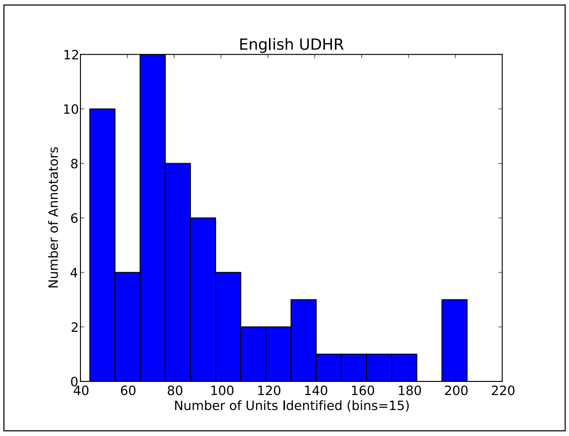

Figure 2: Distribution of total number of annotations per annotator for English UDHR.

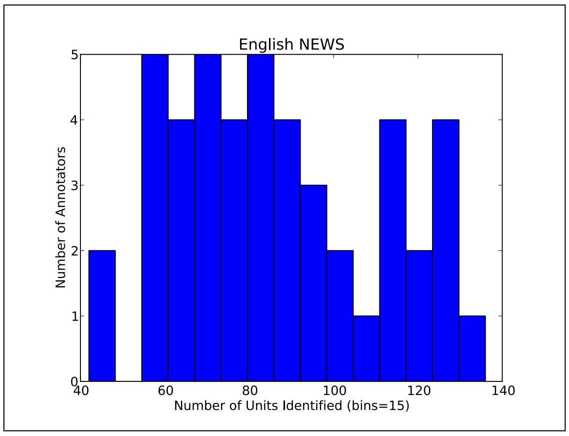

Figure 3: Distribution of total number of annotations per annotator for English NEWS.

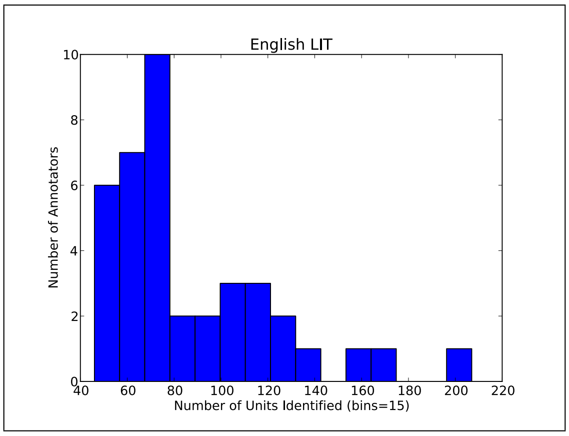

Figure 4: Distribution of total number of annotations per annotator for English LIT.

we used 2-sample Kolmogorov-Smirnov (KS) Tests to identify separate distributions for each of these genres. We found 3 distinct groups in English UDHR (G1–G3) and 2 distinct groups in English LIT (G1 and G2). Table 2 provides more detailed information about distributions for number of annotations, as well as the average number of units found, and average words per unit. This information informs us as to how large or small on average the meaning units are. Note that in Table 2 we include information for overall English UDHR and overall English LIT distributions for reference. The authors found it interesting that, from Table 2, the number of words per meaning unit generally followed the 7 +/- 2 “chunks” phenomenon, where chunks are words.

5.2 Annotator Agreement

Even though some of the annotators agreed about the number of units, that does not imply that they agreed on where the boundaries were placed. We used segmentation similarity (S) as a metric for annotator agreement. The algorithm requires specifying a unit of measurement between boundaries – in our case we used word-level units for English data and character-level units for Chinese data. We calculated average similarity agreement for segment boundaries pair-wise within-group for annotators from each of the 9 language/genre datasets, as presented in Table 3.

While the segmentation similarity agreements seem to indicate high annotator agreement, we wanted to find out if that agreement was better than what we could generate at random, so we compared annotator agreement with random baselines. To generate the baselines, we used the average number of segments per paragraph in each language/genre dataset and inserted boundaries at random. For each of the 9 language/genre datasets, we generated 30 baseline samples. We calculated the baseline segmentation similarity

Dataset	(S)	(SBL)
Chinese UDHR	**0.930**	0.848
English NEWS	**0.891**	0.796
English LIT	**0.875**	0.790
English LIT G1	**0.929**	0.824
English LIT G2	**0.799**	0.727
English UDHR	**0.870**	0.802
English UDHR G1	**0.929**	0.848
English UDHR G2	**0.910**	0.836
English UDHR G3	**0.826**	0.742

Table 3: Within-group segmentation similarity agreement (S) and segmentation similarity agreement for random baseline (SBL).

(SBL) in the same way using average pair-wise agreement within-group for all of the baseline datasets, shown in Table 3.

For English UDHR, we also calculated average pair-wise agreement across groups, shown in Table 4. For example, we compared English UDHR G1 with English UDHR G2, etc. Human annotators consistently outperformed the baseline across groups for English UDHR.

Dataset	(S)	(SBL)
English UDHR G1–G2	**0.916**	0.847
English UDHR G1–G3	**0.853**	0.782
English UDHR G2–G3	**0.857**	0.778

Table 4: English UDHR across-group segmentation similarity agreement (S) and random baseline (SBL).

6 Analysis

Constructing concepts in this task is systematic as was shown from the segmentation similarity scores. Since we know that the annotators agreed on some things, it is important to find out what they have agreed on. In our analysis, we examined unit boundary locations across genres in addition to phrase structure using constituency parses. In this section, we begin to address another of our original research questions regarding how well speakers agree on meaning unit boundary positions across genres and which syntactic features are the most salient for meaning units.

6.1 Unit Boundary Positions for Genres

Boundary positions are interesting because they can potentially indicate if there are salient parts of the texts which stand out to annotators across genres. We have focused this analysis across genres for the overall data for each of the 4 language/genre pairs. Therefore, we have omitted the subgroups – English UDHR groups (G1,G2, G3) and English LIT groups (G1, G2). Although segmentation similarity is greater within-group from Table 3, this was not enough to inform us of which boundaries annotators fully agree on. For each of the datasets, we counted the number of annotators who agreed on a given boundary location and plotted histograms. In these plots we show the number of annotators of each potential boundary between words. We show the resulting distributions in Figures 5 – 8.

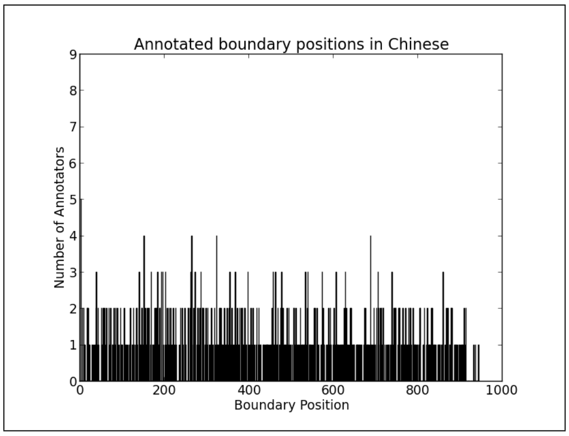

Figure 5: Annotated boundary positions Chinese UDHR.

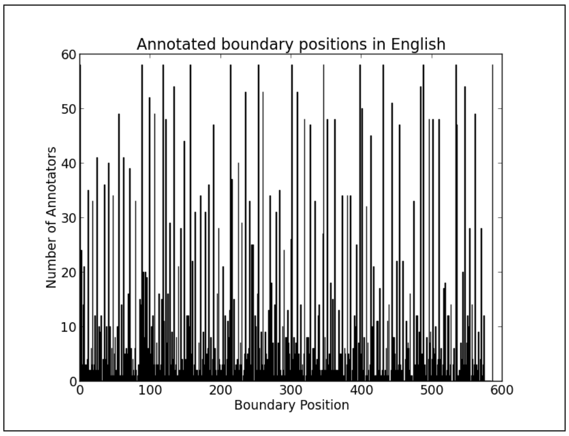

Figure 6: Annotated boundary positions English UDHR.

While there were not many annotators for the Chinese UDHR data, we can see from Figure 5

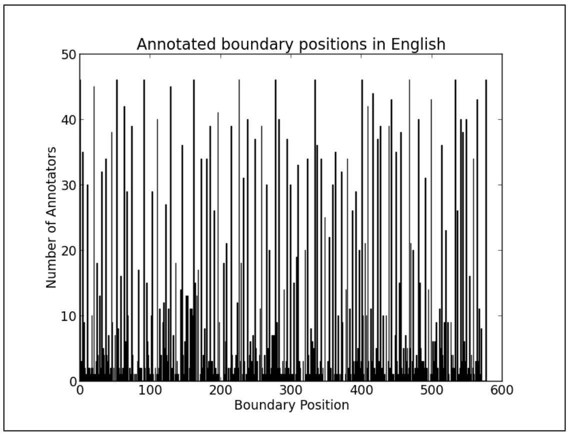

Figure 7: Annotated boundary positions English NEWS.

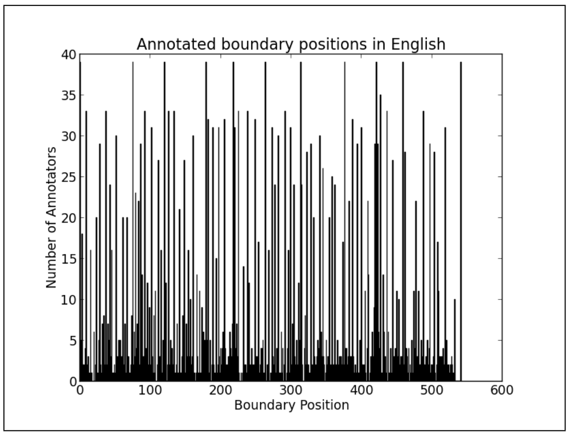

Figure 8: Annotated boundary positions English LIT.

that at most 4 annotators agreed on boundary positions. We can see from Figures 6 – 8 that there is high frequency of agreement in the text which corresponds to paragraph boundaries for the English data, however paragraph boundaries were artificially introduced into the experiment because each paragraph was numbered.

Since we had removed all punctuation markings, including periods and commas for both languages, it is interesting to note there was not full agreement about sentence boundaries. While we did not ask annotators to mark sentence boundaries, we hoped that these would be picked up by the annotators when they were constructing meaning units in the text. Only 3 sentence boundaries were identified by at most 2 Chinese UDHR annotators. On the other hand, all of the sentence boundaries were idenfied for English UDHR and English NEWS, and one sentence boundary was unmarked for English LIT. However, there were no sentence boundaries in the English data that were marked by all annotators - in fact the single most heavily annotated sentence boundary was for English NEWS, where 30% of the annotators marked it. The lack for identifying sentence boundaries could be due to an oversight by annotators, or it could also be indicative of the difficulty and ambiguity of the task.

6.2 Phrase Structure

To answer our question of whether or not there are salient syntactic features for meaning units, we did some analysis with constituency phrase structure and looked at the maximal projections of meaning units. For each of the 3 English genres (UDHR, NEWS, and LIT) we identified boundaries where at least 50% of the annotators agreed. For the Chinese UDHR data, we identified boundaries where at least 30% of annotators agreed. We used the Stanford PCFG Parser on the original English and Chinese text to obtain constituency parses (Klein & Manning, 2003), then aligned the agreeable segment boundaries with the constituency parses. We found the maximal projection corresponding to each annotated unit and we calculated the frequency of each of the maximal projections. The frequencies of part-of-speech for maximal projections are shown in Tables 5 - 8. Note that the part-of-speech tags reflected here come from the Stanford PCFG Parser.

Max. Projection	**Description**	**Freq.**
S, SBAR, SINV	Clause	28
PP	Prepositional Phrase	14
VP	Verb Phrase	11
NP	Noun Phrase	5
ADJP	Adjective Phrase	3
ADVP	Adverb Phrase	1

Table 5: Frequency of maximal projections for English UDHR, for 62 boundaries.

Max. Projection	**Description**	**Freq.**
S, SBAR, SINV	Clause	30
VP	Verb Phrase	23
NP	Noun Phrase	11
PP	Prepositional Phrase	3
ADVP	Adverb Phrase	2

Table 6: Frequency of maximal projections for English NEWS, for 69 boundaries.

Max. Projection	Description	Freq.
S, SBAR	Clause	32
VP	Verb Phrase	10
NP	Noun Phrase	3
PP	Prepositional Phrase	2
ADVP	Adverb Phrase	2

Table 7: Frequency of maximal projections for English LIT, for 49 boundaries.

Max. Projection	Description	Freq.
NN, NR	Noun	22
VP	Verb Phrase	8
NP	Noun Phrase	8
CD	Determiner	3
ADVP	Adverb Phrase	1
AD	Adverb	1
VV	Verb	1
JJ	Other noun mod.	1
DP	Determiner Phrase	1

Table 8: Frequency of maximal projections for Chinese UDHR, for 46 boundaries.

Clauses were by far the most salient boundaries for annotators of English. On the other hand, nouns, noun phrases, and verb phrases were the most frequent for annotators of Chinese. There is some variation across genres for English. This analysis begins to address whether or not it is possible to identify syntactic features of meaning units, however it leaves open another question as to if it is possible to automatically identify a 1-to-1 mapping of concepts across languages.

7 Discussion and Future Work

We have presented an experimental framework for examining how English and Chinese speakers make meaning out of text by asking them to label places that they could construct concepts with as few words as possible. Our results show that there is not a unique “meaning unit” segmentation criteria among annotators. However, there seems to be some preferential trends on how to perform this task, which suggest that any random segmentation is not acceptable. As we have simplified the task of meaning unit identification by using well-structured text from the Universal Declaration of Human Rights, news, and literature, future work should examine identifying meaning units in transcribed speech.

Annotators for the English UDHR and English LIT datasets could be characterized by their different granularities of annotation in terms of number of units identified. These observations are insightful to our first question: what granularity do people use to construct meaning units? For some, meaning units consist of just a few words, whereas for others they consist of longer phrases or possibly clauses. As we did not have enough responses for the Chinese UDHR data, we are unable to comment if identification of meaning units in Chinese fit a similar distribution as with English and we leave in-depth cross-language analysis to future work.

A particularly interesting finding was that human annotators share agreement even across groups, as seen from Table 4. This means that although annotators may not agree on the number of meaning units found, they do share some agreement regarding where in the text they are creating the meaning units. These findings seem to indicate that annotators are creating meaning units systematically regardless of granularity.

Our findings suggest that different people organize and process information differently. This is a very important conclusion for discourse analysis, machine translation and many other applications as this suggests that there is no optimal solution to the segmentation problems considered in these tasks. Future research should focus on better understanding the trends we identified and the observed differences among different genres. While we did not solicit feedback from annotators in this experiment, we believe that it will be important to do so in future work to improve the annotation task. We know that the perceived lag time in speech-to-speech translation cannot be completely eliminated but we are interested in systems that are “fast” enough for humans to have quality conversations in different languages.

Acknowledgments

This work was partly supported by Singapore Agency for Science, Technology and Research (A-STAR) and the Singapore International Pre-Graduate Award (SIPGA) and was partly supported by the National Science Foundation (NSF) award IIS-1225629. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of A-STAR and NSF.

References

Chang Baobao, Pernilla Danielsson, and Wolfgang Teubert. 2002. Extraction of translation units from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing - Volume 18 (SIGHAN '02)*, 1–5.

Presentación Padilla Benítez and Teresa Bajo. 1998. Hacia un modelo de memoria y atención en interpretación simultánea. *Quaderns. Revista de traducció*, 2:107–117.

Chris Fournier and Diana Inkpen. 2012. Segmentation and similarity agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*, Montreal, Canada, 152–161.

Osamu Furuse and Hitashi Iida. 1996. Incremental translation utilizing constituent boundary patterns. In *Proceedings of the 16th conference on Computational linguistics (COLING '96)*, Copenhagen, Denmark, 412–417.

Olivier Hamon, Christian Fgen, Djamel Mostefa, Victoria Arranz1, Munstin Kolss, Alex Waibel, and Khalid Choukri. 2009. End-to-End Evaluation in Simultaneous Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, (EACL '09)*, Athens, Greece, 345–353.

Daniel Jurafsky. 1988. Issues in relating syntax and semantics. In *Proceedings of the 12th International conference on Computational Linguistics (COLING '88)*, Budapest, Hungary, 278–284.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden, 60–67.

Walter Kintsch. 2005. An Overview of Top-down and Bottom–up Effects in Comprehension: The CI Perspective. *Discourse Processes*, 39(2&3):125–128.

Walter Kintsch and Praful Mangalath. 2011. The Construction of Meaning. *Topics in Cognitive Science*, 3:346–370.

Dan Klein and Christopher D. Manning 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.

Jáchym Kolář. 2008. *Automatic Segmentation of Speech into Sentence-like Units.* Ph.D. thesis, University of West Bohemia, Pilsen, Czech Republic.

Patrik Lambert, Adrià. De Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation (LREC)*, 39:267–285.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT '12)*, Montreal, Canada 243–252.

George A. Miller. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity of Processing Information. *The Psychological Review*, Vol 63:81–97.

Hideki Mima, Hitoshi Iida, and Osamu Furuse. 1998. Simultaneous interpretation utilizing example-based incremental transfer. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '98)* Montreal, Quebec, Canada, 855–861.

Pierre Oléron and Hubert Nanpon. 1965. Recherches sur la traduction simultanée. *Journal de Psychologie Normale et Pathologique*, 62(1):73–94.

Mathais Paulik and Alex Waibel. 2009. Automatic Translation from Parallel Speech: Simultaneous Interpretation as MT Training Data. *IEEE Workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, 496–501.

Lev Pevzner and Marti A. Hearst 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):1936. MIT Press, Cambridge, MA, USA.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Mar- tin, and Dan Jurafsky. 2004. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04).*

Baskaran Sankaran, Ajeet Grewal, and Anoop Sarkar. 2010. Incremental Decoding for Phrase-based Statistical Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, Uppsala, Sweden, 222–229.

Teresa M. Signorelli, Henk J. Haarmann, and Loraine K. Obler. 2011. Working memory in simultaneous interpreters: Effects of task and age. *International Journal of Billingualism*, 16(2): 192–212.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, Qun Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing (SIGHAN '03)* - Volume 17, Sapporo, Japan, 184-187.