Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts

Yong Xu, François Yvon

Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts

Yong Xu, François Yvon
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
{yong, yvon}@limsi.fr

Abstract

Resources for evaluating sentence-level and word-level alignment algorithms are unsatisfactory. Regarding sentence alignments, the existing data is too scarce, especially when it comes to difficult bitexts, containing instances of non-literal translations. Regarding word-level alignments, most available hand-aligned data provide a complete annotation at the level of words that is difficult to exploit, for lack of a clear semantics for alignment links. In this study, we propose new methodologies for collecting human judgements on alignment links, which have been used to annotate 4 new data sets, at the sentence and at the word level. These will be released online, with the hope that they will prove useful to evaluate alignment software and quality estimation tools for automatic alignment.

Keywords: Parallel corpora, Sentence Alignments, Word Alignments, Confidence Estimation

1. Introduction

Bitext alignment consists of finding corresponding units in bitexts, where a bitext is defined as the association of two texts assumed to be mutual translations. Such a mapping can be established at various levels of granularity: between paragraphs, between sentences, between phrases, or between words. Primarily because of the development of Statistical Machine Translation (SMT) technologies (Brown et al., 1993), sentence-level and word-level alignments have been studied for a long time. In state-of-the-art phrase-based SMT, sentence alignment aims at providing parallel sentence pairs for word alignment which is an important component of the complete pipeline (Koehn et al., 2003). Their uses extend to many other natural language processing (NLP) applications. For instance, sentence alignment has been applied in translator training (Simard et al., 1993), translation checking (Macklovitch, 1994), language learning (Nerbonne, 2000; Kraif and Tutin, 2011), and bilingual reading (Pillias and Cubaud, 2015). Word alignment is employed in bilingual lexica extraction (Smadja et al., 1996), word sense disambiguation (Diab and Resnik, 2002), etc. Thanks to a sustained research effort, many alignment methods have been proposed. Two recent reviews of bitext alignment are in (Wu, 2010; Tiedemann, 2011).

Manually annotated reference alignment data sets are valuable resources for the development of alignment techniques. On the one hand, they can be used as the supervision examples for the methods (Mújdricza-Maydt et al., 2013; Blunsom and Cohn, 2006); on the other hand, they provide ways to directly evaluate automatic alignment quality, and warrant the investigation of error patterns. However, constructing manually annotated alignment data sets can be challenging. For some tasks, this can be due to a lack of a clear annotation scheme. For others, annotation schemes can vary a lot, depending on the targeted applications, language pairs, etc.

In this paper, we describe our contribution to manual sentence-level alignment annotations in Section 2., followed by word-level alignment annotations in Section 3.. For sentence alignment, the research community has reached a consensus on the annotation scheme (Tiedemann, 2011). But the resource is quite scarce for certain types of bitexts. We report, in § 2.1., our collection of manual sentence alignments for literary bitexts, a challenging usecase for alignment techniques. Next, in § 2.2., we propose a new scheme for annotating parallel fragments, which has been used to label data set of candidate parallel sentences. These resources might prove useful for tasks such as confidence estimation, or for filtering incorrect pairs in a translation memory. Regarding word alignments, our view is to consider one-to-one and many-to-many links separately. We present a novel set of annotation labels for one-to-one links in § 3.2., and a collection of annotations using these tags. We then describe an innovative methodology for collecting many-to-many word alignment links in § 3.3., as well as the corresponding data set.

All the data sets described in this paper, except the first one, were created by three annotators pursuing a master level degree in translation studies, who were retributed for this work. Two of them are native French speakers with advanced capacities in English and Spanish. The other annotator is a native Greek speaker, fluent in English and French. For each task, the annotators were given guidelines, and applied them to annotate a small amount of sandbox instances (which are not included in the final data sets). Potential ambiguities regarding the task and the guidelines were then discussed and resolved, in order to a) ensure a shared understanding of the principles and details, and b) if necessary, improve the guidelines. In a second step, the actual data sets were annotated.

2. Sentence alignments

2.1. Reference alignments for literary works

Given a bitext $E_1^I = E_1, ..., E_I$ (source side) and $F_1^J = F_1, ..., F_J$ (target side), where each E_i or F_j is a sentence, sentence alignment is the task of recovering sentence-level alignment links between the two sides, i.e. finding the corresponding sentence groups. An alignment *link* has two sides, each containing any number (including 0) of con-

Book	Language pair	# Link	# Source sent.	# Target sent.
Du Côté de chez Swann (M. Proust)	EN-FR	463	495	492
Emma (J. Austen)	EN-FR	164	216	160
Jane Eyre (C. Brontë)	EN-FR	174	205	229
La Faute de l'Abbe Mouret (E. Zola)	EN-FR	222	226	258
Les Confessions (J.-J Rousseau)	EN-FR	213	236	326
Les Travailleurs de la Mer (V. Hugo)	EN-FR	359	389	405
The Last of the Mohicans (F. Cooper)	EN-FR	197	205	232
* Alice's Adventures in Wonderland (L. Carroll)	EN-FR	746	836	941
* Candide (Voltaire)	EN-FR	1,230	1,524	1,346
* Hound of the Baskervilles (A. Conan Doyle)	EN-FR	822	862	893
* Vingt Mille Lieues sous les Mers (J. Verne)	EN-FR	778	820	781
* Voyage au Centre de la Terre (J. Verne)	EN-FR	714	821	754
* Candide (Voltaire)	EN-EL	1,247	1,524	1,585
* Candide (Voltaire)	EN-ES	1,113	1,524	1,196
Total 14 books		8,442	9,883	9,598

Table 1: Statistics of reference sentence alignments for literary works. We use the terms "source" and "target" for convenience only, as they do not indicate the actual original language. All "source" entries refer to English. Alignments marked with a * are refinements of A. Farkas' initial alignments. The others are revised version of the data presented in (Yu et al., 2012).

secutive sentences.[1] For example, $[E_i, E_{i+1}; F_j]$ denotes a 2-to-1 link. Sentence alignment is a helpful processing step in many NLP applications, such as SMT.

Compared to words and phrases, sentences in bitexts typically exhibit a higher level of translational regularity: sentences are generally translated in monotonic order; in some types of bitexts, like technical manuals, most sentences are translated one-by-one. According to these observations, the research community has reached the following assumptions for computing sentence-level alignments (Tiedemann, 2011) :

- Each side of an alignment link is a consecutive group of sentences, or is empty. That is, if E_i and E_{i+2} are both inside a link, then so must be E_{i+1}.
- Links must be minimal, in the sense that they cannot be decomposed into strictly smaller links. For example, if both $[E_i; F_j]$ and $[E_{i+1}; F_{j+1}]$ are good alignment links, then it is incorrect to form a larger link $[E_i, E_{i+1}; F_j, F_{j+1}]$.
- Alignment links are monotone. Thus, if $[E_i; F_j]$ is a link, then no source sentences *following* E_i (e.g. E_{i+1}) can link to target sentences *preceding* F_j (e.g. F_{j-1}).

A main advantage of these assumptions is that they warrant the use of dynamic programming to perform efficient search. To our knowledge, all automatic sentence alignment systems make such assumptions.

Classical sentence alignment systems were initially designed to align institutional bitexts (Brown et al., 1991; Gale and Church, 1991), such as the Canadian Hansards and the Europarl corpus (Koehn, 2005). The ARCADE evaluation campaigns (Véronis and Langlais, 2000; Chiao et al., 2006) have demonstrated that the quality of automatic alignments is variable, depending on the bitext genres and languages. For certain types of bitexts which are relatively regular, such as institutional bitexts, the task is easy and all systems tend to deliver good results (the basic system of Brown et al. (1991) obtained above 95% precision on the Hansards). On the contrary, for literary bitexts, alignment quality could be much less satisfactory. Yu et al. (2012) and Lamraoui and Langlais (2013) reported that the best link-level F-score obtained for "De la Terre à La Lune" (J. Verne), a part of the BAF corpus (Simard, 1998), was only around 78%. Hence, literary bitexts, which typically include larger portions of non literal translations, would be very useful to evaluate the actual performance of state-of-the-art alignment systems. To our knowledge, however, there are few publicly available reference sentence alignments for literary works, the most used being the BAF corpus. The need for gold alignments for such materials has also been pointed out in (Yu et al., 2012; Lamraoui and Langlais, 2013).

To alleviate this scarce resource problem, we have collected manual alignments for a small set of literary works. Our annotators have processed excerpts from 12 classical books for French-English. Smaller Greek-English and Spanish-English corpora have also been collected, notably resulting in a mutiple sentence alignment of Voltaire's "Candide". The annotation was performed using the Uplug toolkit (Tiedemann, 2003). In order to make our annotations more suited to evaluate automatic alignment tools, the annotators have made sure that our manual alignments actually follow the conventions listed above (minimality, monotonicity, prohibition of gappy alignments, etc).

Table 1 summarizes the main statistics of the corpus. Note that for the books with the * mark, alignment links were generated as refinements of existing reference paragraph

[1] If one side of a link is empty, it is called a null link. However a 0-to-0 link makes no sense and is not allowed.

alignments provided by A. Farkas.[2]
We do not report agreement figures here because the task is relatively easy and well understood. In a sandbox experiment, the agreement rate between three annotators is as high as 99.8%.

2.2. Confidence in sentence alignment

Automatic alignments are mostly used for statistical machine translation (Koehn, 2005). In this context, it is custom to filter out unreliable alignment links based on heuristic confidence estimation measures, such as length ratios. Confidence estimation can also prove useful in other contexts, for instance in bilingual concordancers (Simard et al., 1993; Bourdaillet et al., 2009) for translator training or in other language learning scenarios. This is even more necessary when alignments are extracted from noisy bitexts, e.g. bitexts collected from the internet (Tiedemann, 2011), or for crowd-sourced alignments.
Confidence Estimation (CE) for sentence alignments aims at judging the usability of alignment links. This is different from quality estimation for machine translation, where the quality of system outputs as valid sentences is not assured and plays an important role. In CE for sentence alignment, all sentences are deemed to be well formed, and the only thing that needs to be evaluated is the level of correspondence between the two sides of a link. However, the usability of a link depends on the targeted application. The canonical sentence alignment evaluation metric, the F-measure, distinguishes two classes (correct and wrong). The recently introduced task of translation memory (TM) checking considers three cases:[3] a pair of segments can be a) totally correct and need no editing at all, or b) need substantial editing, or c) can be mostly correct but need few simple edits. Similar categories have been used to assess the usefulness of automatic translations for post-edition (Wisniewski et al., 2013). Note finally that for SMT, even partially correct alignments and very loose mutual translations can be useful as training data.
To better reflect this flexible notion of alignment link quality, we propose a new annotation scheme based on a 5-way categorization of sentence alignment links:

1. *sure*: the pair of sentences are (near) perfect mutual translations;
2. *partial*: one side contains some parts that are not translated on the other side.
3. *unperfect*: the pair constitutes a loose complete mutual translation, or a translation only in specific contexts;
4. *erroneous*: the pair has no correspondence relation at all (i.e. the pair is not correct);
5. *undecidable*: none of the previous four cases, where corresponds to cases where for pairs which are highly context-dependent and cannot be annotated in isolation. In practice, this class is quite rarely used.

Note that upon choosing the label "partial", our annotators were also asked to mark explicitly the untranslated part (see 1).
We took the automatically sentence-aligned English-French OpenSubtitle Corpus,[4] randomly picked 1,800 alignment links, and used the proposed tags to annotate them. Each alignment link was annotated twice. The first annotator annotated links $3000 \sim 4199$ (link ID range), the second annotated $3600 \sim 4799$, and the third annotated $4200 \sim 4799$ and $3000 \sim 3599$. We used an adapted version of the Yawat tool (Germann, 2008) to perform this task. Figure 1 displays the annotation process. We found that the inter-annotator agreement for this task was very high (the average $\kappa \approx 0.85$), showing that our annotation scheme was sensible. Among the 1,663 links that the annotators agreed on the labels, 1,002 were tagged as "sure" (62.25%), 252 "partial" (15.15%), 163 "unperfect" (9.80%), 244 "erroneous" (14.67%), and 2 "undecidable" (0.12%).

3. Collecting subsentential alignment information: two new proposals

3.1. Evaluating word alignments with gold references

Bilingual word alignments constitute an important resource for many downstream applications in multilingual NLP. Some rely on 1-to-1 alignment links, e.g. in cross-lingual transfer of Part-of-Speech labels (Täckström et al., 2013; Wisniewski et al., 2014) or of other kinds of information; others use many-to-many alignments, e.g. phrase-based SMT (Koehn et al., 2003). Most applications perform better when alignment quality is improved (Lambert et al., 2005). Because word alignment is both important and challenging, it has received a sustained attention of the research community since the introduction of IBM Models by Brown et al. (1993). Numerous approaches have been since proposed to improve alignment quality ((Liang et al., 2006; Dyer et al., 2011; Wang et al., 2015), to name a few).

Metrics The evaluation of word alignments is, however, a tricky question (Tiedemann, 2011). On the one hand, compared to sentence alignments, word alignments suffer from much more severe ambiguity problems. It is often difficult, if possible at all, for annotators to agree on the correctness of certain alignment links. On the other hand, the notion of alignment quality can only be understood in reference to some targeted application. Applications such as bilingual lexical extraction prefer high precision word alignments, while others such as SMT might prefer high recall alignments (Och and Ney, 2004). Therefore, evaluation of word alignments typically include both intrinsic and extrinsic metrics. The most commonly used intrinsic evaluation metric for word alignment is the Alignment Error Rate (AER) proposed by Och and Ney (2000). It relies on a particular annotation scheme for gold alignments, which distinguishes between **S**ure links and **P**ossible links. AER amounts to a F1 measure where recall and precision are computed differently for these two types of links. This metric and the corresponding annotation scheme have been

[2] http://FarkasTranslations.com

[3] This scheme is used for the shared task on cleaning translation memories of the NLP4TM'16 workshop. See http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/.

[4] Downloadable from http://opus.lingfil.uu.se/.

Figure 1: Sentence alignment confidence annotation. For each alignment link, the color of the special symbol "$$$" encodes its label: green for "sure", violet for "partial", etc. Note the untranslated part of the pair 3003 (labelled "partial") appears in gray.

criticized in many subsequent studies (Fraser and Marcu, 2007), notably due to the lack of clear semantics of **P**-links, which tend to be used in too many situations (non-literal translation, many-to-many alignments, etc.). Regarding extrinsic metrics, a widely used approach is to consider SMT output quality measured by automatic scores such as BLEU. As repeatedly noted (Lopez and Resnik, 2006; Fraser and Marcu, 2007; Lambert et al., 2010), AER poorly correlates with translation quality, especially for large corpora, which makes the direct comparison of alignment systems more difficult.

Building reference alignments The construction of gold word alignments is a complicated task: their specification must address deep linguistic issues (which are often specific to language pairs), but also take into account the intended use of these alignments, notwithtanding more concrete issues such as interface design and disagreement resolution procedures. Melamed (1998) was the first to propose a complete annotation guideline for the Blinker project, which was used to align 250 verse pairs of the Bible (English-French) with a binary annotation scheme. Och and Ney (2000) used the Blinker guidelines to align 484 sentence pairs of the Hansard corpus (English-French), further introducing the Sure/Possible distinction. Mihalcea and Pedersen (2003) collected a set of English-Romanian word alignments for 265 sentence pairs, again using the Blinker guidelines and the S/P scheme. Lambert et al. (2005) created guidelines to align 500 sentence pairs of the English-Spanish version of Europarl, with the explicit purpose to create high recall alignments. Some more recent works are (Kruijff-Korbayova et al., 2006) (English-Czech), (Graça et al., 2008) (multiple language pairs), (Macken, 2010) (English-Dutch), (Holmqvist and Ahrenberg, 2011) (English-Swedish), etc, most of them sticking to the S/P scheme.

We propose new methodologies to collect evaluation data for word alignment. Our proposal relies on two distinct protocols: the first focuses on 1-to-1 alignments and proposes on a much clarified version of the S/P distinction (see § 3.2.); the second specifically targets many-to-many alignments, and is based on a divisive annotation strategy which proceeds iteratively (see § 3.3.). For both tasks, the annotations are carried out with adapted versions of Yawat.

3.2. A new annotation scheme for 1-to-1 alignments

The S/P annotation scheme was designed for one-to-one alignment links. One major problem with this scheme is the vagueness of this distinction, yielding annotations that are highly subjective. In (Och and Ney, 2000), it is stated that: "a S (sure) alignment which is used for alignments which are unambiguously and a P (possible) alignment which is used for alignments which might or might not exist". Yet, for some annotators, an unambiguous link might imply a context-independent word pair; for others, if a source word A is in the context of a particular sentence pair the best match for target word B, and vice-versa, then the link is unambiguous. Many-to-one alignments are also often difficult to annotate. Second, the vagueness of **P** links makes their systematic exploitation difficult: for instance, when a multiword expression is paraphrased, it is common practice to **P**-tag all individual word links in the corresponding block (Lambert et al., 2005; Graça et al., 2008). This block of **P** links would be helpful for a multiword expression extractor; however, some other **P** links are made of word pairs that share the same meaning in a particular context and that would be irrelevant for such an application. Lambert et al. (2005) further pointed out that reference alignments having a large majority of **P** links would limitate the usefulness of the AER metric, as automatic alignments of very different underlying quality might achieve the same AER score with respect to such a reference dataset.[5]

We hold the view that, for annotations to be maximally useful, the **S** tag should indicate word pairs that can reliably used in any application, thus it should be reserved for word pairs that share the same meaning in most contexts (a similar semantics for the **S** tag is used in (Graça et al., 2008)). As for **P** links, we find that the majority of them fall into two categories: some are contextual, while others are part of a larger correspondence between *groups of words*. We thus propose to define the following annotation tags for 1-to-1 word alignment links:

- *sure*: the pair of words express the same meaning, e.g. "dog – chien";
- *contextual*: the pair of words express the same mean-

[5] When a group of source words are aligned to a group of target words, it is custom to **P**-tag all resulting 1-to-1 links in the Cartesian product. Unfortunately, this can easily lead to a large number of 1-to-1 **P** links in the reference.

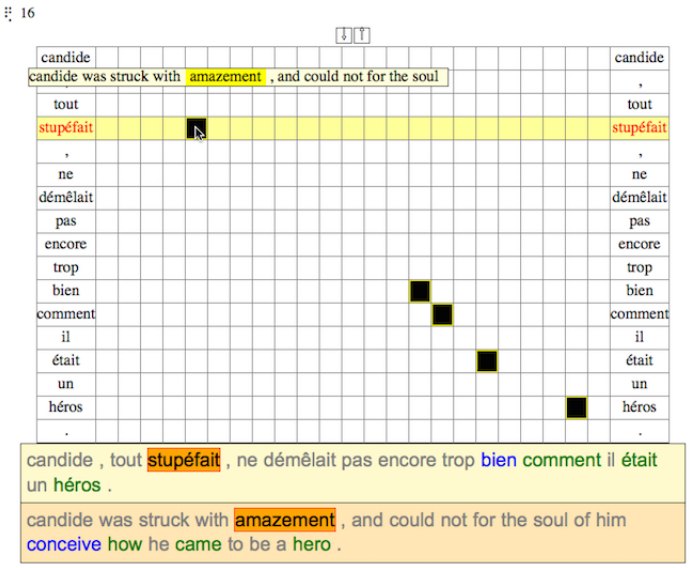

Figure 2: 1-to-1 word alignment confidence annotation for a parallel sentence from "Candide". A black cell in the alignment matrix represents a potential alignment link. For each link, the color of the word pair corresponds to its label.

ing only in the specific context, e.g. "tomorrow – samedi" (French for "Saturday");

- *partial*: the pair of words do not constitute a good link by themselves, but they should be included in a larger link (group of words), e.g. "(make) use (of) – (se) servir (de)";
- *wrong*: the corresponding pair of words should not be aligned.

This annotation scheme has been tested using high-confidence 1-to-1 links produced automatically. This set of alignments was prepared as follows. For each language pair, we first combined the sentence-aligned "Candide" and the Europarl data for this language pair (Koehn, 2005) into a parallel corpus, which was word-aligned by running MGIZA (Gao and Vogel, 2008) in both directions. We then formed a small candidate corpus, by taking all sentence pairs of "Candide" and a few hundreds of the Europarl.[6] Finally, for each sentence pair in the candidate corpus, we have selected at most five 1-to-1 links in the intersection of the directional alignments, thereby ensuring that the potential alignment points were sensible choices.

Each link was then manually annotated with one of the four labels described above. Figure 2 illustrates the annotation process for one parallel sentence from "Candide" (French-English). Using this methodology, we were able to collect 2,691 link annotations for English-French, 3,118 for English-Spanish, 2,996 for Spanish-French, 2,204 for Greek-English, and 527 for Greek-French, totaling 11,536 word-level annotations. On the English-French subset of links that were hand-annotated more than once, the inter-annotator agreement rate is around 0.75. Figure 3 shows the distribution of labels per language pair.[7] We observe that

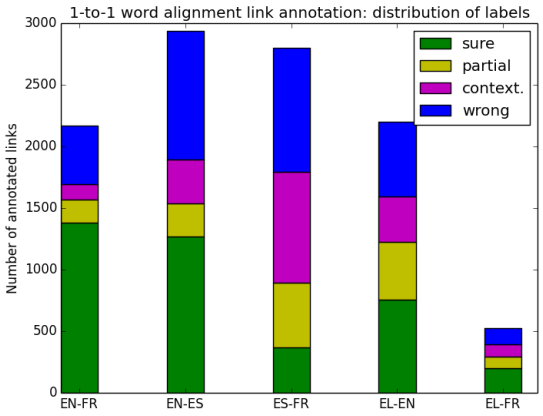

Figure 3: The distribution of 1-to-1 word alignment annotation labels per language pair.

each of the labels "partial" and "contextual", though less frequently used than "sure" and "wrong" in general, represents a non-negligible, sometimes even important, portion. This observation confirms our belief that a finer categorization than **S**ure and **P**ossible is sensible. The distribution of labels varies for each language pair. The most remarkable situation is perhaps the large proportion of links tagged as "contextual" in the Spanish-French data, which certainly requires further study.

3.3. Collecting reference many-to-many alignments

We further propose a novel method to obtain reference many-to-many alignments. The protocol is based on recursive divisions of parallel sentence pairs. Given a pair of parallel segments (we call such a pair of segments a *bi-*

[6] We ran MGIZA on the combined large corpus instead of just "Candide" to maximise the quality of automatic word alignments.

[7] For some subsets annotated by more than one annotator, we have taken the intersection. So the numbers of links in Figure 2 are slightly different from those reported in the text.

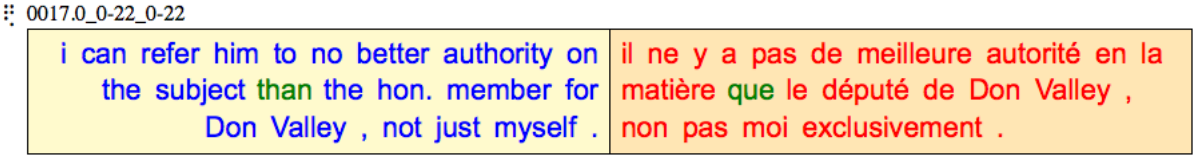

Figure 4: The first pass, splitting sentence pair 0017.

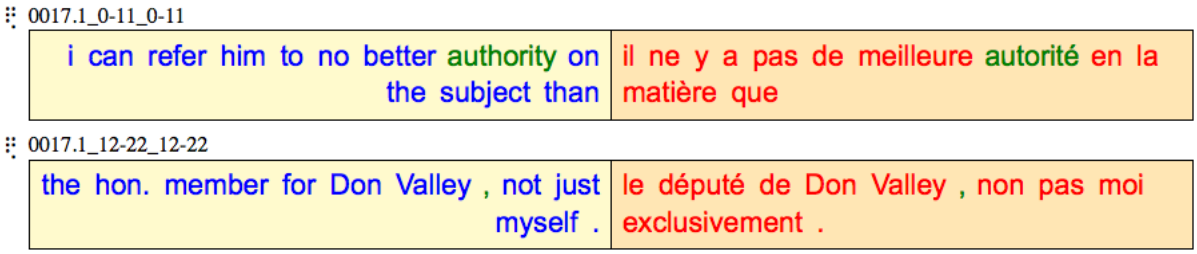

Figure 5: The second pass, splitting the two bi-segments of sentence pair 0017.

segment) E_1^I with I words and F_1^J with J words, the annotators iterated the following process:

1. If the bi-segment cannot be further divided, terminate;
2. Else, pick an index i for E, an index j for F, such that the four segments $E_1^i, E_{i+1}^I, F_1^j, F_{j+1}^J$ can form two bi-segments. One possibility is that E_1^i is parallel with F_1^j and E_{i+1}^I is parallel with F_{j+1}^J; another is that E_1^i is parallel with F_{j+1}^J and E_{i+1}^I is parallel with F_1^j (the indices i and j define *splitting points*);
3. For each bi-segment produced in step 2, go to step 1.

We believe that this protocol is much simpler than annotating the full alignment matrix, since at each step there is only one single decision to make. Two heuristics are used to guide the annotation process: (a) when many segmentation index pairs (i and j in step 2) are acceptable splitting points, choose one such that (i) the segmentation is as balanced as possible (in terms of segment lengths), (ii) the linguistic structures are preserved as much as possible; (b) the process terminates when either the bi-segment in step 1 is a 1-to-1 word pair, or when the segment is not strictly compositional and thus cannot be split.

Figure 4 illustrates the first iteration of splitting a sentence pair. Note many splitting points other than the chosen one are possible: for example, the pair of words "authority" and "autorité". But the resulting two bi-segments would be less balanced. We may even choose "on" and "en", but this would destroy the pair of expressions (though compositional) "on the subject" and "en la matière", which we prefer to keep together. Figure 5 displays the second iteration, during which we split the two resulting bi-segments of the first pass. It is obvious that we can continue to proceed in this way, for instance, by processing the bi-segment "the hon. member for Don Valley" and "le député de Don Valley".

We have recorded all the bi-segments generated during the whole process, resulting in a hierarchical alignment structure between the original sentences. Ideally, for a parallel sentence, all annotators would arrive at the same set of final bi-segments (albeit with different choices of splitting points). This is hard to achieve in reality, since annotators might differ on the notion of linguistic structures and compositionality. Still, the bi-segment sets can help to estimate our confidence for many-to-many alignments. If a computed many-to-many alignment can be decomposed into a combination of several bi-segments, then it is reasonable to suggest that it is a good link. Furthermore, the hierarchical nature of our annotations makes it possible to design metrics for word alignments that could explicitly depend on their compatibility with our bi-segmentation.

For this data set, the annotators were presented with 1,086 sentence pairs from Europarl, 220 from the Hansard, and 290 from Jules Verne's "Vingt Mille Lieues sous les Mers" and Sir Arthur Conan Doyle's "The Great Shadow". The final set contains approximately 10,000 bi-segments.

The 220 Hansard sentence pairs were chosen from the trial and test set of the NAACL 2003 workshop on word alignment (Mihalcea and Pedersen, 2003), where reference word alignment had been provided by Och and Ney (2000). This subset enables us to compare our minimal bi-segments with this reference alignment. For these 220 sentences, our recursive segmentation method gave rise to 3,971 final bi-segments, which contained 2,540 (64.0%) one-to-one links, 451 (11.4%) two-to-one links, 133 (3.3%) three-to-one, and 335 (8.4%) links whose both sides had more than 2 words (many-to-many). The reference alignment of these 220 sentence pairs contains 2,720 **S** one-to-one links, among which only 37 are not included in any of our bi-segments. In other words, our bi-segmentations contain a large majority of the **S** one-to-one links of Och and Ney (2000). We believe this partially confirms the value of our annotation scheme. Comparatively, the analysis of **P** one-to-one links is much less satisfactory, since only 4,328 (out of 9,915) **P** one-to-one links in the reference of 2003 are actually included in one of our bi-segments, demonstrating again the uncertainty of these correspondences.

4. Conclusion

In this paper, we have described several data sets, all designed for the purpose of evaluating bitext alignments softwares with a special attention to their possible use for confidence estimation purposes. We have analyzed the alignment annotation tasks and discussed the weaker points of existing annotation schemes. Based on this analysis, we have proposed new annotation schemes for both sentence and word level alignments. We contribute also a method for collecting reference many-to-many alignments, which, we believe, is an innovative attempt for direct evaluation of this kind of alignments. The resources and corresponding annotation guidelines are publicly available.[8]

We plan to use these annotations to evaluate results delivered by standard sentence alignment and word alignment tools. In particular, we are interested in using these data to evaluate confidence estimation measures, e.g. based on posterior link probabilities (Huang, 2009).

Another lesson learned in this annotation exercices is that sentence-level and word-level alignments are quite sensitive to the pre-processing, e.g. sentence segmentation, tokenization in words, etc. It might be beneficial to investigate new ways to overcome these man-made noises so as to produce gold annotations that would be less dependent on these early steps.

5. Acknowledgements

This work was partly supported by French National Research Agency under project Transread (ANR-12-CORD-0015). We thank L. Berenice, C. Clément and M. Sgourelli for performing the annotations. We have made good use of the alignment data from ©2014 FarkasTranslations.com and would thus like to thank András Farkas for making his multi-parallel corpus of manually aligned books publicly available.

6. Bibliographical References

Blunsom, P. and Cohn, T. (2006). Discriminative word alignment with conditional random fields. In *Proceedings of ACL*, pages 65–72.

Bourdaillet, J., Huet, S., Gotti, F., Lapalme, G., and Langlais, P. (2009). Enhancing the bilingual concordancer transsearch with word-level alignment. In *Proceedings of LNAI*, pages 27–38.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of ACL*, pages 169–176.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chiao, Y.-C., Kraif, O., Laurent, D., Nguyen, T. M. H., Semmar, N., Stuck, F., Véronis, J., and Zaghouani, W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of LREC*.

[8] https://transread.limsi.fr/resources.html.

Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL*, pages 255–262.

Dyer, C., Clark, J. H., Lavie, A., and Smith, N. A. (2011). Unsupervised word alignment with arbitrary features. In *Proceedings of ACL*, pages 409–419.

Fraser, A. and Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of ACL*, pages 177–184.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Germann, U. (2008). Yawat: Yet Another Word Alignment Tool. In *Proceedings of the ACL-08: HLT Demo Session*, pages 20–23.

Graça, J. a., Pardal, J. P., Coheur, L., and Caseiro, D. (2008). Building a golden collection of parallel multi-language word alignment. In *Proceedings of LREC*.

Holmqvist, M. and Ahrenberg, L. (2011). A gold standard for English-Swedish word alignment. In *Proceedings of NODALIDA*, pages 106–113.

Huang, F. (2009). Confidence measure for word alignment. In *Proceedings of ACL-IJCNLP*, pages 932–940.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of NAACL: HLT*, pages 48–54.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT summit*, pages 79–86.

Kraif, O. and Tutin, A. (2011). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. In *Corpora, Language, Teaching, and Resources: From Theory to Practice. Selected papers from TaLC7*.

Kruijff-Korbayova, I., Chvatalova, K., and Postolache, O. (2006). Annotation guidelines for Czech-English word alignment. In *Proceedings of LREC*, pages 1256–1261.

Lambert, P., De Gispert, A., Banchs, R., and Mariño, J. B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Lambert, P., Petitrenaud, S., Ma, Y., and Way, A. (2010). Statistical analysis of alignment characteristics for phrase-based machine translation. In *Proceedings of EAMT*.

Lamraoui, F. and Langlais, P. (2013). Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment? In *Proceedings of MT Summit*, pages 77–84.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of NAACL*, pages 104–111.

Lopez, A. and Resnik, P. (2006). Word-based alignment, phrase-based translation: What's the link. In *Proceedings of AMTA*, pages 90–99.

Macken, L. (2010). An annotation scheme and gold stan-

dard for Dutch-English word alignment. In *Proceedings of LREC*.

Macklovitch, E. (1994). Using bi-textual alignment for translation validation: the TransCheck system. In *Proceedings of AMTA*, pages 157–168.

Melamed, I. D. (1998). Annotation style guide for the blinker project. Technical report, Dept. of Computer and Information Science, University of Pennsylvania.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, pages 1–10.

Mújdricza-Maydt, E., Köerkel-Qu, H., Riezler, S., and Padó, S. (2013). High-precision sentence alignment by bootstrapping from wood standard annotations. *The Prague Bulletin of Mathematical Linguistics*, (99):5–16.

Nerbonne, J., (2000). *Parallel Texts in Computer-Assisted Language Learning*, chapter 15, pages 354–369. Text Speech and Language Technology Series.

Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of COLING*, pages 1086–1090.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Pillias, C. and Cubaud, P., (2015). *Proceedings of INTERACT 2015*, chapter Bilingual Reading Experiences: What They Could Be and How to Design for Them, pages 531–549.

Simard, M., Foster, G., and Perrault, F. (1993). Transsearch: A bilingual concordance tool. Technical report, Centre for Information Technology Innovation.

Simard, M. (1998). The BAF: a corpus of English-French bitext. In *Proceedings of LREC*, pages 489–494.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Täckström, O., Das, D., Petrov, S., Ryan, M., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. In *Transactions of the ACL*.

Tiedemann, J. (2003). *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Proceedingsssing*. Ph.D. thesis, Uppsala University.

Tiedemann, J. (2011). *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies.

Véronis, J. and Langlais, P. (2000). Evaluation of Parallel Text Alignment Systems. In *Parallel Text Processing*, Text Speech and Language Technology Series, chapter X, pages 369–388.

Wang, X., Utiyama, M., Finch, A., Watanabe, T., and Sumita, E. (2015). Leave-one-out word alignment without garbage collector effects. In *Proceedings of EMNLP*, pages 1817–1827.

Wisniewski, G., Singh, A. K., and Yvon, F. (2013). Quality estimation for machine translation: Some lessons learned. *Machine Translation*, 27(3).

Wisniewski, G., Pécheux, N., Gahbiche-Braham, S., and Yvon, F. (2014). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of EMNLP*, pages 1779–1785.

Wu, D. (2010). Alignment. In *CRC Handbook of Natural Language Processing*, number 16, pages 367–408.

Yu, Q., Max, A., and Yvon, F. (2012). Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *Proceedings of BUCC*.