Chapter 7

Examining lexical coherence in a multilingual setting

Karin Sim Smith
The University of Sheffield

Lucia Specia
The University of Sheffield

This paper presents a preliminary study of lexical coherence and cohesion in the context of multiple languages. We explore two entity-based frameworks in a multilingual setting in an attempt to understand how lexical coherence is realised across different languages. These frameworks (an entity-grid model and an entity graph metric) have previously been used for assessing coherence in a monolingual setting. We apply them to a multilingual setting for the first time, assessing whether entity based coherence frameworks could help ensure lexical coherence in a Machine Translation context.

1 Introduction

We present an exploratory study which represents our early research on how lexical coherence is realised in a multilingual context, with a view to identifying patterns that could be later used to improve overall translation quality in Machine Translation (MT) models.

Ideally a coherent source document when translated properly should result in a coherent target document. Coherence does vary in how it is achieved in different languages. Moreover, unlike a human translator, who translates the document as a whole, in context, ensuring that the translated document is as coherent as the source document, most MT systems, and particularly Statistical

Karin Sim Smith & Lucia Specia. 2017. Examining lexical coherence in a multilingual setting. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 119–137. Berlin: Language Science Press. DOI:10.5281/zenodo.814470

Machine Translation (SMT) systems, translate each sentence in isolation, and have no notion of discourse principles such as coherence and cohesion.

While some research has indicated that MT frameworks are good at lexical cohesion (Carpuat & Simard 2012), in that they are consistent, others have reported different results (Wong & Kit 2012), since MT systems can persist using with a particular translation which is incorrect. We believe that investigating entity-based frameworks in a multilingual setting may shed some light on the issue. In particular, we also hope to ascertain whether they help in the disambiguation of lexical entities, where in an MT setting the translation of a particular source word, e.g. 'bank' in English, could be translated as either 'la rive' or 'la banque' in French, depending on the context. Currently most SMT systems determine which word to use purely based on the probabilities established at training time (i.e. how frequently 'bank' equated to 'la rive' and how frequently it equated to 'la banque'). While, this should be determined by context, the problem is that most systems translate one sentence at a time, disregarding the wider context.

Greater insight into how multilingual lexical coherence is achieved could lead to improvements in current translation approaches. This improvement could take the form of features based on the entity transitions, guiding the lexical choice. Alternatively, we could use coherence models to select the option which leads to a higher translation score when reranking results from a decoder.

In the following (Section 2) we describe entity based coherence. We briefly explain the grid model (Section 3) and the graph one (Section 4). Then we detail our experimental settings (Section 5) for the two main parts of this research. Firstly (Section 6), we constructed a multilingual comparative entity-based GRID for a corpus comprising various documents covering three different languages. We examine whether similar patterns of entity transitions are exhibited, or whether they varied markedly across languages. Secondly (Section 7), we applied an entity **graph** in a multilingual context, using the same corpus. We assess whether this different perspective offers more insight into crosslingual coherence patterns. Our conclusions are set out in Section 8. Our goals are to understand differences in lexical coherence across languages so that in the future we can establish whether this can be used as a means of ensuring that the appropriate level of lexical coherence is transferred from source to machine translated documents.

2 Entity-based coherence

There has been recent work in the area of lexical cohesion in MT (Xiong et al. 2013a,b; Tiedemann 2010; Hardmeier 2012; Carpuat & Simard 2012; Wong & Kit

2012), as a sub category of coherence, looking at the linguistic elements which hold a text together. However, there seems to be little work in the wider area of coherence as a whole. Coherence is indeed a more complex discourse element to define in the first place. While it does include cohesion, it is wider in terms of also describing how a text becomes semantically meaningful overall, and how easy it is for the reader to follow.

Xiong et al. (2013b) focus on ensuring lexical cohesion by reinforcing the choice of lexical items during decoding. They subsequently compute lexical chains in the source text (Xiong et al. 2013a), project these onto the target text, and integrate these into the decoding process with different strategies. This is to try and ensure that the lexical cohesion, as represented through the choice of lexical items, is transferred from the source to target text. Tiedemann (2010) attempts to improve lexical consistency and to adapt statistical models to be more linguistically sensitive, integrating contextual dependencies by means of a dynamic cache model. Hardmeier (2012) suggests there is not much to be gained by just enforcing consistent vocabulary choice in SMT, since the vocabulary is already fairly consistent. While there is indeed a case for arguing that MT systems can be more consistent than human translators for using a set terminology (Carpuat & Simard 2012), that would only be valid for a very narrow field, perhaps a highly technical domain, and an SMT system trained on exact data. Wong & Kit (2012) study lexical cohesion as a means of evaluating the quality of MT output at document level, but in their case the focus is on it as an evaluation metric. Their research supports the intuition we found, i.e. that human translators intuitively ensure cohesion, which in MT output often is represented as direct translations of source text items that may be inappropriate in the target context. They conclude that MT needs to learn to use lexical cohesion devices appropriately.

Lexical cohesion is only one aspect of coherence, however much of the work on computationally determining how lexical cohesion is indicative of coherence refers to 'coherence', therefore we retain the term 'coherence' here, as we are looking at how lexical cohesion contributes to coherence as a whole. In particular, the focus, or the 'attentional state' (Grosz & Sidner 1986) in a discourse is one major aspect of coherence. Entity-based coherence aims to measure the attentional state, formalised via Centering Theory (Grosz, Weinstein & Joshi 1995) (more below).

The entity-based approach was first proposed by Barzilay & Lapata (2005) with the aim of measuring local coherence in a monolingual setting, focusing on applications where multiple alternatives of a system output are available, such as the ranking of alternative automatic text summaries by their coherence degree.

As detailed by Barzilay & Lapata (2008), the entity-based approach derives from the theory that entities in a coherent text are distributed in a certain manner, as identified in various discourse theories, in particular in Centering Theory (Grosz, Weinstein & Joshi 1995). This theory holds that coherent texts are characterised by salient entities in strong grammatical roles, such as subject or object. The focus of their work (Barzilay & Lapata 2008) was in using this knowledge, via patterns in terms of prominent syntactic constructions, to distinguish coherent from non-coherent texts. In our research the focus is on differences in the general patterns, particularly across languages. As long as a syntactic parser is available, this approach is fully automatic and avoids human annotation effort. We see it as a means of extracting additional linguistic information for use in rich features to guide lexical selection in MT, as well as potentially in the problem of MT evaluation.

Previous computational models for assessing coherence have been deployed in a monolingual setting (Lapata 2005; Barzilay & Lapata 2008; Elsner, Austerweil & Charniak 2007; Elsner & Charniak 2011; Burstein, Tetreault & Andreyev 2010; Guinaudeau & Strube 2013). We report on our findings for applying the entity grid (Section 6) and entity graph (Section 7) to a multilingual setting, using data and settings as described in Section 5.

Our initial experiments will take all nouns in the document as discourse entities, as recommended by Elsner & Charniak (2011), and investigate how they are realised crosslingually. The distribution of entities over sentences may vary from language to language (more on this below). The challenge from an MT point of view would be to ensure that an entity chain is carried over to from source to target text, despite differences in syntax and sentence structure, and taking account of linguistic variations.

3 Entity grid

Entity distribution patterns vary according to text domain, style and genre, which are all valuable characteristics to capture, and attempt to transfer from source to target text languages where appropriate. They are constructed by identifying the discourse entities in the documents under consideration and representing them in 2D grids whereby each column corresponds to the entity, i.e. noun, being tracked, and each row represents a particular sentence in the document in order. An example can be seen in Table 1, where the lines represent consecutive sentences, and the columns ('e1', etc.) represent different entities. In this example, 'e7' represents 'Kosovo', which was repeated in sentences 's2', 's3' and 's4', in the roles of *subject* (S), *other* (X), and *subject* (S), respectively.

Table 1: Example of entity grid

	e1	e2	e3	e4	e5	e6	e7
s1	-	-	-	-	-	-	-
s2	-	-	-	-	-	-	S
s3	-	-	-	-	-	-	X
s4	-	-	O	-	-	-	S
s5	S	-	-	-	-	-	-
s6	-	-	-	X	-	-	-

Once all occurrences of nouns and the syntactic roles they represent in each sentence (Subject (S), Object (O), or other (X)) are extracted, an ENTITY TRANSITION is defined as a consecutive occurrence of an entity, with given syntactic roles. These are computed by examining the grid vertically for each entity. For example, an 'SS', a 'Subject-to-Subject' transition, indicates that an entity occurs in a subject position in two consecutive sentences. An 'SO', on the other hand, indicates that while the entity was in a subject role in one sentence, it became the object in the subsequent sentence. Probabilities for these transitions can be easily derived by calculating the frequency of a particular transition divided by the total number of transitions which occur in that document.

4 Entity graph

Guinaudeau & Strube (2013) projected the entity grid into a graph format, using a bipartite graph which they claim had the advantage both of avoiding the data sparsity issues encountered by Barzilay & Lapata (2008) and of achieving equal performance on measuring overall document coherence without the need for training. They use it to capture the same entity transition information as the entity grid experiment, although they only track the occurrence of entities, avoiding the nulls or absences of the other (tracked as '-' in the entity grid framework). Additionally, the graph representation can track cross-sentential references, instead of only those in adjacent sentences (Guinaudeau & Strube 2013).

The graph tracks the presence of all entities, taking all nouns in the document as discourse entities, as recommended by Elsner & Charniak (2011), and connections to the sentences they occur in. The general form of the coherence score assigned to a document in this approach is shown in Equation 7.1. This is a centrality measure based on the average outdegree across the N sentences represented in the document graph. The outdegree of a sentence s_i, denoted $o(s_i)$, is the total weight leaving that sentence, a notion of how connected (or how cen-

tral) it is. This weight is the sum of the contributions of all edges connecting s_i to any $s_j \in D$.

$$s(D) = \frac{1}{N} \sum_{i=1}^{N} o(s_i) \tag{7.1}$$
$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} W_{i,j}$$

The coherence of a text in this model is measured by calculating the average outdegree of a projection, so by summing the shared edges (i.e. of entities leaving a sentence) between two sentences.

They define three types of graph projections: *binary*, *weighted* and *syntactic*. Binary projections simply record whether two sentences have any entities in common. Weighted projections take the number of shared entities into account, rating the projections higher for more shared entities. A syntactic projection includes syntax information, where syntactic information is used to weight the importance of the link by calculating an entity in role of subject (S) as a 3, an entity in role of object (O) as a 2, and other (X) as a 1. These are projected between any two sentences in the text, as sets of shared entities.

We projected the entity relationships onto a graph-based representation, as per Guinaudeau & Strube (2013), experimenting in various settings. Our objective was to assess whether the graph gives us a better appreciation of differences in entity-based coherence across languages. This representation can encode more information than the entity-grid as it spans connections not just between adjacent sentences, but among all sentences in the document.

5 Experimental settings

For our multilingual experiments, the entity grid approach was applied to parallel texts from the WMT corpus,[1] with three languages: English, French, and German. In particular, we used the test data, comprising news excerpts extracted over various years. The direction of translation varies for different documents, as discussed in Section 6. For comparison, we also take the French and English documents from the **LIG** corpus (Potet et al. 2012) of French into English translations. These form a concatenated group of 361 documents, which are news

[1] http://www.statmt.org/wmt10/

excerpts drawn from various WMT years. In all these corpora, translations are provided by human, professional translators.

French to English is generally regarded as a well performing language pair in MT, whereas German to English is more error-prone due to compounding, word order and morphological variations in German. Of particular interest here are the compound words prevalent in German, and how these affect the entity grid. To establish general tendencies, entity grids were compiled for three different sources:

- The **newstest2008** datasets in each language comprising 90 parallel documents.
- The **LIG** corpus in French and English comprising 361 parallel documents.

In our experiments we used version 3.3.0 of the Stanford Parser[2] to identify the noun phrases in each language. We set the salience at 2, i.e. recording only entities which occurred more than twice, and derived models with transitions of length 3 (i.e. over 3 adjacent sentences). We computed the mean of the transition probabilities, i.e. the probability of a particular transition occurring, over all the documents.

While previous work for English, a language with a relatively fixed word order, has found factors such as the grammatical roles associated with the entities affect local coherence, this varies across languages (Cheung & Penn 2010). Cheung & Penn (2010) further suggest that topological fields (identifying clausal structure in terms of the positions of different constituents) are an alternative to grammatical roles in local coherence modelling, for languages such as German, and show that they are superior to grammatical roles in an ordering experiment.

For this set of experiments we therefore apply a slightly simplified version of the grid, recording the presence or absence of particular (salient) entities over a sequence of sentences. In addition to being the first cross-lingual study of the grid approach, this experiment also aims at examining the robustness of this approach without a syntactic parser. While the grammatical function may have been useful as an indicator in the aforementioned work, this does not necessarily hold in a multilingual context. Simply tracking the existence or absence of entities allows for direct comparison across languages. Indeed, as Filippova & Strube (2007) reported when applying the entity grid approach to group related entities and incorporate semantic relatedness, "syntactic information turned out to have a negative impact on the results". While Barzilay & Lapata (2008) argued that "the proposed representation reveals entity transition patterns characteristic of coherent texts", we would also suggest that these patterns potentially vary

[2] http://nlp.stanford.edu/software/corenlp.shtml

from language to language to some extent, while retaining an overall degree of coherence.

6 Multilingual grids

6.1 In-depth analysis

In order to illustrate the differences between the distributions of entity transitions over the different languages, we computed Jensen-Shannon divergence scores for French and English, and then German and English, both displayed in Figure 1.

Paying attention to the scale, it is clear that the German and English divergence is greater overall than the divergence for French and English. For example the entity transitions which showed the highest variation were $XX-$, which was 0.045 for the difference between French and English over 0.1 for German and English, also transition XXX where the difference over the same was 0.02 and 0.08. This indicates that for the German-English pair it was less likely that the same entity showed up in 3 consecutive sentences than for the French-English pair.

Table 2: Multilingual entity transitions (mean of 90 documents)

Transition	German	French	English
'XXX'	0.001445	0.002382	0.000441
'X-X'	0.006240	0.006917	0.003184
'XX-'	0.005905	0.008853	0.003130
'-XX'	0.004142	0.006155	0.001672

While German is more nominal in structure, and one might expect higher entity transition probabilities in general, these are often compound nouns, which are then counted separately in our setup. This variance merits more investigation to gain a fuller picture of the reasons behind it.

There is a clear pattern across the entity transitions over the three languages studied. In this instance we are comparing the same texts, on a document by document basis, so the same genre and style, yet there is a consistent difference in the probabilities. This would appear to indicate, amongst other things, that the manner in which lexical coherence is achieved varies from language to language.

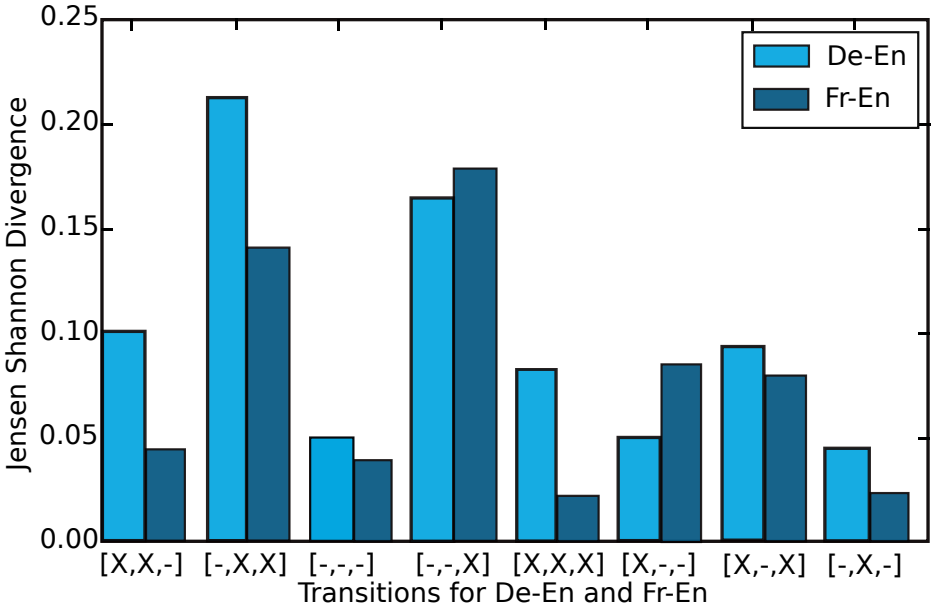

Figure 1: Jensen-Shannon divergence over distribution of entity transitions (length 3) for German-English and French-English (WMT newstest2008)

While this is just a preliminary study with a small dataset, this is supported by other research findings (Lapshinova-Koltunski 2015b).

On closer analysis, it would appear that there are various issues at play. Firstly, there is the matter of sentence boundaries, which affects the transition probabilities. Across many of the documents in the **newstest2008**, the French version had fewer sentences within segments than the corresponding segments in German or English. This potentially increases the number of transitions from sentence to sentence. French also exhibited on average fewer entities per document. So the transitions are more concentrated. Both of these factors potentially account for some of the higher levels of entity transitions in French over English and German in the WMT **newstest2008** documents.

The tendency in the WMT **newstest2008** documents was for English and German to have more, shorter sentences. So elements of discourse which were in one sentence in French were occasionally split over two sentences in German or English, and thus an entity transition was over two consecutive sentences in French, but had a sentence between them in the other two languages. As a re-

sult, the XXX transition count was typically higher for French. Interestingly, French also exhibited a higher count of $XX-$ transitions, often over sentences 1 and 2. Of course, we can enforce the constraint of strictly parallel sentences, but it is interesting to see the natural linguistic variation.

6.2 Linguistic trends

Interestingly, another reason for the variation across languages may be the fact that in French there is a tendency to use a noun in the plural as well as singular. For example, in document 37 of the **LIG** corpus the French used 2 separate entities where the English had one: 'inequality', which occurred at positions: 0, 1, 2, 3, 4, 12, 13, 14, 17, 18, 19, 21, 31, was rendered in French by 2 separate entities: 'inégalités' at 0, 1, 2, 4, 12, 14, 17, 18, 19, 31 'l'inégalité' at 2, 3, 13.

This phenomenon occurred elsewhere too: 'effort' in English occurred in the following sentences of document 24: 8, 9, 10, 11. In French we actually find 3 separate entities used, due to the way the parser dealt with the definite article: 'l'effort' at 8, 'effort' at 9, 11 and 'efforts' at 9, 10. While we can adapt our models (via lemmatisation) to account for the linguistic variation, it is important that we appreciate the linguistic variation in the first place, if we are to ensure **appropriate** lexical coherence.

In addition, sometimes an entity in English is actually rendered as an adjective in French, and therefore not tracked in the grid, such as document 5, where the source text, i.e. French, has 'crises cambiaires' rendered in the English as 'currency crises', and while 'currency' is identified as an entity in English, it is an adjective in French, thus not identified as an entity. Apart from affecting the transition probabilities, it would seem that some form of lexical chains is necessary to fully capture all the necessary lexical information in this multilingual setting. In the same document, 'currency' occurs 8 times as an entity in the English, yet in the French besides being rendered as an adjective twice, is rendered 4 times as 'caisse d'émission' and only once as 'monnaie'. This is reflected in the fact that for this document the English had 127 entities where the French had 152.

Another interesting point to note is that in general German exhibited a higher entity count. This is to be expected, as German is more nominal in structure than, for example, French. This count is also affected by the amount of compound verbs in German, and how we decide to model them. Thus, for example, from a document on cars, the word 'car' features as a main entity, but whereas it appears 4 and 6 times in French ['voiture' at sentences 6, 8, 23, 31, 32, 33] and English ['car' at sentences 5, 7, 22, 31, 32, 33] respectively, in German it only appears twice ['Auto' at sentences 7, 22]. However, 'car' is part of a collection of compound

words in German, such as 'High-end-auto' at sentence 31 in the document, [31=X] and 'Luxusauto' at sentence [32=X]. As it occurs in a different form, it is, in this instance, tracked as a different entity altogether.

Similarly, German exhibited a high ratio of $X - X$ transitions, where an entity skips a sentence, then reoccurs. This is explained by the occurrence of more, shorter sentences, as described above, and also by the compounding factor. With shorter sentences there is a greater chance that entities are split between two sentences, where the French may have had one. This also leads to lower likelihood of a transition to the next sentence; the transition would instead skip one sentence (appear as $X - X$ transition instead of $XX-$ or XXX). Plus a particular entity may not appear in three consecutive sentences, as it may have done in the French or English versions, because in the middle sentence it is part of a compound verb.

This illustrates the linguistic differences that need to be taken into account when examining comparative coherence in a multilingual context. This could lead to a decision to lemmatise before extracting grids or graphs, but in that case they are no longer strictly **entity** grids. We can apply linguistic processing to make the different grids comparable, but that should be sensitive to the linguistic variation, as overly processing to make them comparable will lose the natural expression in a particular language.

6.3 Source language implications

In some cases the quality of the text was also an issue. WMT data (from which the **LIG** corpus was also derived) is generated both from texts originally in a given language, e.g. English, and texts manually translated from other languages (e.g. Czech) into that language (say English). And in some cases the human translation of the documents was not particularly good. This was the case for some of the English documents translated from Czech in the **newstest2008** corpus. This has a direct influence on the coherence of the text, yet as noted by Cartoni et al. (2011), often those using this WMT corpus fail to realise the significance of whether a text is an original or a translation.

What also has to be taken into account is the language of the source text, and the tendency for it to affect the target text in style, depending on how literal the translation is.

6.4 Entity realisations

It is interesting to trace how the main entities in a given text are realised across the languages. See Table 3 where each numbered column represents a sentence

in that parallel document. We have cut the last few sentences from the table, in order to fit it in.

Table 3: Occurrences of 'Brown' in various sentences of parallel document (dropping last sentences of document due to spacing)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
DE	x	-	-	x	x	x	-	-	-	-	-	x	-	-	-	x	-	x
FR	x	-	-	x	x	x	-	-	-	-	-	x	-	-	-	x	-	x
EN	x	-	x	-	x	-	-	-	-	-	x	-	-	-	x	-	x	-
	18	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
DE	-	x	-	x	-	-	-	-	-	x	-	-	-	-	x	x	-	-
FR	-	x	-	-	x	-	-	-	-	-	x	-	-	-	-	x	x	-
EN	x	-	-	x	-	-	-	-	-	x	-	-	-	-		x	-	-

We can clearly see how the main subject is realised through the document, albeit not at identical positions. On occasion, this is affected by differences in sentence breaks. In this case the French and German entities were closely matched in position at the start of the document, and then the English and German by the end. However, the point is that in general, there are the same number of occurrences, as the thread of discourse is traced through each document with exact positions dependent on sentence breaks. This pattern of occurrences is valuable information which among other things can potentially be used to improve anaphora resolution in the target text. Centering Theory has been used (Kehler 1997) to resolve referents by working out the backward looking centre for a sentence. Thus one of the entities referred to in one sentence may well be referred to in a subsequent sentence by a reference (Clarke & Lapata 2010). This study in entity grids has the potential to be useful in this domain too.

7 Multilingual graphs

7.1 Compound splitting

We also analyse the graph framework in a multilingual setting to try and garner additional insight into variations in coherence patterns in different languages. The intuition is that this framework could be more informative than the grid as it spans connections between not just adjacent sentences, but any subsequent ones.

Our initial experiments take all nouns in the document as discourse entities, as recommended by Elsner & Charniak (2011), and investigate how the projections

are realised by lexical items. As discovered during experiments for the entity grid, the entity spread over sentences may vary from language to language (more on this below).

We used the weighted projection, which considers the frequencies of the various entities in the documents, which we determined was more appropriate than syntax in a comparative multilingual context. As regards incorporating syntax for other models, Strube & Hahn (1999) suggests that for freer word-order languages, "We claim that grammatical role criteria should be replaced by criteria that reflect the functional information structure of the utterances". This is particularly relevant for German. Our intuition is that the weighted projection gives the best appreciation of the cohesive links between sentences, as it gives a higher weighting where they are more frequent, unlike the unweighted one which simply logs the sentences which an entity occurs in.

We used the same WMT dataset as for the grid experiments. The graph coherence scores were computed for all parallel multilingual documents and results are displayed in Table 4.

Table 4: Number of documents (out of 90) for a given language which scored the highest among the 3 languages

	coherence score	**coherence score without compound splitter**
French	26	30
English	47	56
German	17	4

On closer analysis we encountered the same issue with German compounds as for the grid, whereby the entities in the German grid were more sparse, and more discontinuous in nature, due to the fact that compound words accounted for several entities. To establish just how much difference this was making, we also tried applying a compound splitter for German[3]. So for a given entity, we check if it decomposes into several entities, and if so each is entered separately in the graph. This resulted in a more uniform coherence score over the 3 languages. Whereas German had the highest coherence score for only 4 out of the 90 documents when no compound splitter was applied, this figure rose to 17 with a compound splitter. This is perhaps more meaningful when doing crosslingual comparisons.

[3] http://www.danielnaber.de/jwordsplitter, Licensed under the Apache License

7.2 Crosslingual similarity

Interestingly, looking at the coherence scores for all 3 languages, they exhibit remarkably similar graph profiles (Figure 2). As in the documents which result in a low score for English are similarly low for French and German. So it would seem that it is possible to assess lexical coherence as judged by this metric in a crosslingual manner, albeit as one aspect of coherence, not as sufficient to alone judge the overall coherence of the document. As Tanskanen (2006) point out, "cohesion may not work in absolutely identical ways in all languages, but the strategies of forming cohesive relations seem to display considerable similarity across languages".

The English documents had the largest proportion of high coherence scores, scoring highest more often than French or German. This could be a general characteristic that English involves more coherence as expressed via simple entity-based coherence and that in German coherence is possibly achieved through other means. Lapshinova-Koltunski (2015a) illustrate, that languages tend to vary in the way they use discourse features.

It certainly supports our findings in the grid experiments, where English had the highest number of entity transitions. From this it would seem that out of these three languages, German exhibits the least entity based coherence, while the highest scores are exhibited by English, followed by French. As Wong & Kit (2012) note, the lexical cohesion devices have to not only be recognised, but used appropriately. And this may differ from the source text to the target text.

7.3 Source language implications

As mentioned already, it is important for this data set to realise what the source language is, and this is marked up on the documents within the WMT data set. This is relevant because it indicates which languages are original texts and which are translations. The first 30 documents are originally Hungarian and Czech (so documents 0-29 in our code). The subsequent 15 ones are originally French (docs 30-44), the next 15 are Spanish (45-59), the next 15 are English (60-74) then German (75-90). This is interesting, as we can then see patterns emerging of naturally coherent texts. It also means that for a number of documents, our French, German and English versions are all translations. One point to note is that ideally this should be extended over an additional corpus, to gain more data, as otherwise we just have 15 texts of each original language. In the meantime, we can see from Table 5 how these affect the scores assigned under this metric. While it is tempting to consider whether having an original German text means that the

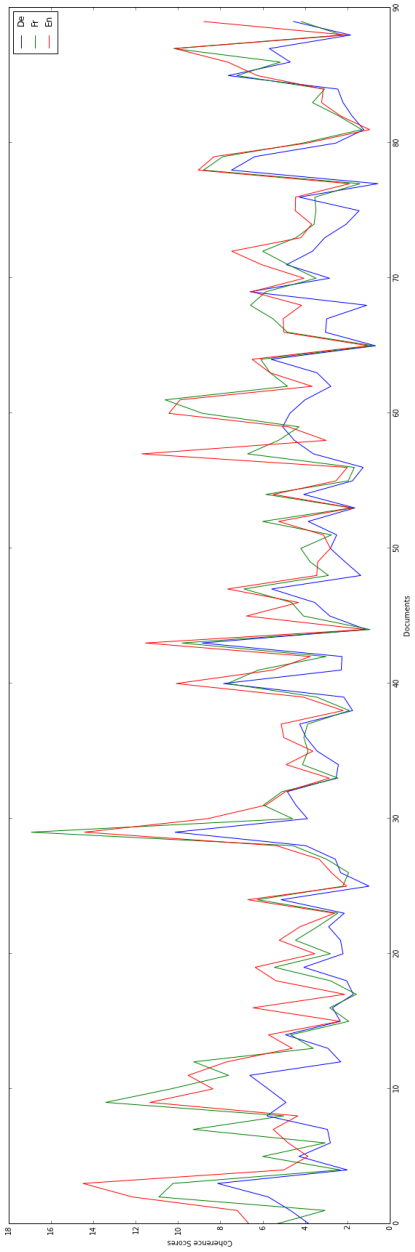

Figure 2: Multilingual graph coherence scores, displaying the score (y-axis) for each document (x-axis)

coherence is higher for German and more evenly scoring in general, or whether an English source text results in less coherence for the German, the number of documents in this preliminary work are not representative enough. This could be worthwhile pursuing as a corpus study, however.

Table 5: Breakdown of highest scoring documents

	French highest	**English highest**	**German highest**
French original (docs 30-44)	3	8	4
English original (docs 60-74)	6	8	1
German original (docs 75-90)	4	6	5

Although the projection score is normalised in that the sum of projections is multiplied by 1/N where N is the number of sentences, there is an inevitable bias in favour of longer documents, for example, document 65 in our experiment using the WMT data has only 3 sentences, and reads as a coherent one, yet due to the shortness has a low score.

Yet document 29, by comparison, scores a high score yet reads incoherently - it is originally Czech, and the translation is clumsy in parts. The high score is due to repetition of words like 'millions', 'krona' or 'year' or their equivalent in French and German. French scores the highest, but seems to also be poor quality.

7.4 Lexical coherence

Intuitively, it would seem that this different perspective, i.e. the graph model, offers more insight into crosslingual coherence patterns, in that it captures all the connections between entities throughout the entire document.

8 Conclusions

We observed distinct patterns in a comparative multilingual approach: the probabilities for different types of entity grid transitions varied, and were generally lower in French than English, with German behind the two, indicating a different coherence structure in the different languages.

The standard format of the grid does, however, need to be modified for a multilingual context. It is clear that there are divergences between languages, as regards entity based coherence. As before, French will still have multiple representations for what would potentially be one entity in English: the use of singular

and plural forms of the noun as noticed in French, or adjectival forms representing the entity. We have also detected differences in implementation due to the compound structure of German; in German while compound nouns affect the coherence score considerably, even with a compound splitter (as for the graph) the coherence score is still generally lower. Possible extensions to this research include expanding the grid to include lexical chains, in place of simple entities, or incorporating a vector of similar terms which would potentially take account of these issues and allow for crosslingual variance in the semantic coverage of an individual lexical item. This would potentially better account for the compound structure of German, and the use of singular and plural forms of the noun as noticed in French, or adjectival forms representing the entity. It is valuable to register and identify the differences and bear them in mind for future development, particularly for crosslingual transfer.

We have seen that the graph leads to a clear picture of entity-based coherence scores. This is perhaps more useful than the grid for comparative studies. We can also see better how entity-based coherence is achieved in different languages. Here the exact sentence breaks do not matter so much, and the score is based on how cohesive the document is as a whole. In future research we will note the significance of whether a text is an original or a translation, filtering our data based on the original language.

Our next step will be to use the graph metric as part of the reranking process within an MT system, to try and assess its ability to disambiguate entities.

The challenge from an MT point of view would be to ensure that the correspondences are maintained, so an entity chain is carried over from source to target text, despite differences in syntax and sentence structure. However, this is insufficient to ensure that the document is fully coherent – more linguistically based elements are necessary to do that.

References

Barzilay, Regina & Mirella Lapata. 2005. Modeling local coherence: An Entity-Based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 141–148. Ann Arbor, Michigan. http://www.aclweb.org/anthology/P05-1018. DOI:10.3115/1219840.1219858

Barzilay, Regina & Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1). 1–34. DOI:10.1162/coli.2008.34.1.1

Burstein, Jill, Joel R. Tetreault & Slava Andreyev. 2010. Using Entity-Based features to model coherence in student essays. In *Proceedings of HLT-NAACL*, 681–684.

Carpuat, Marine & Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of WMT*, 442–449. Montreal, Canada.

Cartoni, Bruno, Sandrine Zufferey, Thomas Meyer & Andrei Popescu-Belis. 2011. How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web* (BUCC '11), 78–86. Stroudsburg, PA, USA: Association for Computational Linguistics.

Cheung, Jackie Chi Kit & Gerald Penn. 2010. Entity-based local coherence modelling using topological fields. In *ACL*, 186–195. http://www.aclweb.org/anthology/P10-1020.

Clarke, James & Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics* 36(3). 411–441.

Elsner, Micha, Joseph Austerweil & Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of HLT-NAACL*, 436–443.

Elsner, Micha & Eugene Charniak. 2011. Extending the entity grid with Entity-Specific features. In *Proceedings of ACL*, 125–129.

Filippova, Katja & Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the 11th European Workshop on Natural Language Generation* (ENLG '07), 139–142. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1610163.1610187.

Grosz, Barbara J. & Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3). 175–204. http://dl.acm.org/citation.cfm?id=12457.12458.

Grosz, Barbara J., Scott Weinstein & Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21. 203–225.

Guinaudeau, Camille & Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of ACL*, 93–103.

Hardmeier, Christian. 2012. Discourse in statistical machine translation. *Discours* (11).

Kehler, Andrew. 1997. Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics* 23(3). 467–475.

Lapata, Mirella. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of IJCAI*, 1085–1090.

Lapshinova-Koltunski, Ekaterina. 2015a. Exploration of inter- and intralingual variation of discourse phenomena. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, 158–167. Lisbon, Portugal.

Lapshinova-Koltunski, Ekaterina. 2015b. Variation in translation: Evidence from corpora. In Claudio Fantinuoli & Federico Zanettin (eds.), *New directions in corpus-based translation studies* (Translation and Multilingual Natural Language Processing), 93–113. Berlin: Language Science Press.

Potet, Marion, Emmanuelle Esperança-Rodier, Laurent Besacier & Hervé Blanchon. 2012. *Collection of a Large Database of French-English SMT Output Corrections*.

Strube, Michael & Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics* 25(3). 309–344. http://dl.acm.org/citation.cfm?id=973321.973328.

Tanskanen, Sanna-Kaisa. 2006. *Collaborating towards coherence: Lexical cohesion in English discourse* (Pragmatics & Beyond New Series). Amsterdam: John Benjamins Publishing Company.

Tiedemann, Jörg. 2010. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, 8–15. Uppsala, Sweden.

Wong, Billy Tak-Ming & Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of EMNLP-CoNLL*, 1060–1068.

Xiong, Deyi, Yang Ding, Min Zhang & Chew Lim Tan. 2013a. Lexical chain based cohesion models for Document-Level statistical machine translation. In *Proceedings of EMNLP*, 1563–1573.

Xiong, Deyi, Guosheng Ben, Min Zhang, Yajuan Lv & Qun Liu. 2013b. Modeling lexical cohesion for Document-Level machine translation. In *Proceedings of IJCAI*.