Generalizing Sub-sentential Paraphrase Acquisition across Original Signal Type of Text Pairs

Aurélien Max **Houda Bouamor** **Anne Vilnat**

LIMSI-CNRS & Univ. Paris Sud
Orsay, France
firstname.lastname@limsi.fr

Abstract

This paper describes a study on the impact of the original signal (text, speech, visual scene, event) of a text pair on the task of both manual and automatic sub-sentential paraphrase acquisition. A corpus of 2,500 annotated sentences in English and French is described, and performance on this corpus is reported for an efficient system combination exploiting a large set of features for paraphrase recognition. A detailed quantified typology of sub-sentential paraphrases found in our corpus types is given.

1 Introduction

Sub-sentential paraphrases can be acquired from text pairs expressing the same meaning (Madnani and Dorr, 2010). If the semantic similarity of a text pair has a direct impact on the quality of the acquired paraphrases, it has, to our knowledge, never been shown what impact the type of original signal has on paraphrase acquisition. In this work, we consider four types of corpora, which we think are representative of the main types of original semantic signals: text pairs (roughly, sentences) originating *a*) from independent translations of a text (TEXT), *b*) from independent translations of a speech (SPEECH), *c*) from independent descriptions of a visual scene (SCENE), and *d*) from independent descriptions of some event (EVENT). We will report the results of experiments on sub-sentential paraphrase acquisition on all these corpus types in two languages, English and French, and provide some answers to the following questions: What types of paraphrases can be found by human annotators, with what confidence and in which quantities? How well can representative paraphrase acquisition systems perform on each corpus type, and how performance can be improved through combination? On what corpus types can performance be improved by using training material from other corpus types? Our experimental results will provide several indications of the differences and complementarities of the corpus types under study, and will notably show that performance on the most readily available corpus type can be improved by using training data from the set of all other corpus types.

We will first describe the building procedures and characteristics of our corpora (section 2), and then describe our experimental settings for evaluating paraphrase acquisition (section 3.1). Our experiments will first consist of the description (section 3.2) and evaluation (section 3.3) of a system combination on each corpus type and then of our system provided with additional training data from the other corpus types (section 3.4). We will finally briefly review related work (section 4) and discuss our main findings and future work (section 5).

2 Collection of sentence pair corpora

In this study, we will focus on paraphrase acquisition from related sentence pairs characteristic of 4 corpus types, which correspond to different original signal types of text pairs illustrated by the word alignment matrices on Figure 1. A corpus for each type has been collected for 2 languages, English and French, and comprises 625 sentence pairs per language. We now briefly describe how each corpus was built.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 721–731, Jeju Island, Korea, 12–14 July 2012.

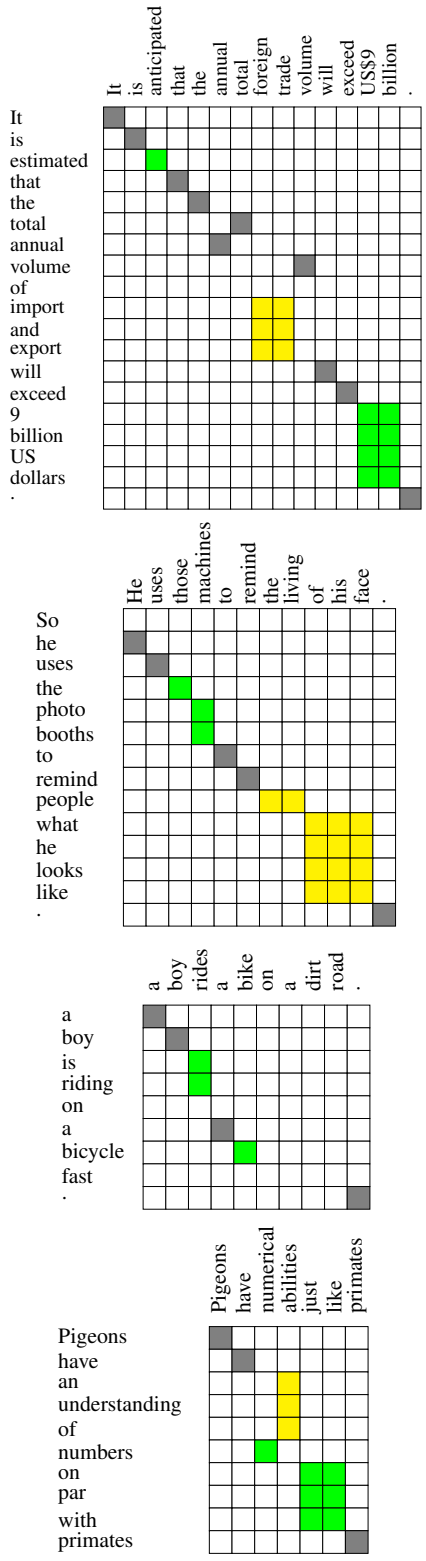

Figure 1: Example reference alignment matrices for (from top to bottom) TEXT, SPEECH, SCENE and EVENT. *Sure* alignments appear in green or gray (identities) and *possible* alignments in yellow.

TEXT For English, we used the MTC corpus[1] (described in (Cohn et al., 2008)) consisting of sets of news article translations from Chinese, and for French the CESTA corpus[2] consisting of sets of news article translations from English. For each sentence cluster, we selected sentence pairs with minimal edit distance above an empirically-selected threshold, covering all clusters first and then selecting from already used clusters to reach the target number of sentence pairs.

e.g. *It is estimated that the total annual volume of import and export will exceed 9 billion US dollars.* ↔ *It is anticipated that the annual total foreign trade volume will exceed US$9 billion.*

SPEECH For English, we used two freely available subtitle files[3] of the French movies *Le Fabuleux Destin d'Amélie Poulain* and *Les Choristes*, and for French we used two subtitle files from the *Desperate Housewives* TV series. We first aligned each parallel corpus using the algorithm described in (Tiedemann, 2007), based on time frames and developed for bilingual subtitles, we then filtered out sentence pairs below a minimal edit distance threshold, and manually removed obvious errors made by the algorithm.

e.g. *So he uses the photo booths to remind people what he looks like.* ↔ *He uses those machines to remind the living of his face.*

SCENE We used the Multiple Video Description Corpus (Chen and Dolan, 2011) obtained from multiple descriptions of short videos. Similarly to what we did for TEXT, we selected sentence pairs from clusters by minimal edit distance above a threshold. An important fact is that for English we were able to use what is described as "verified" descriptions. There were, however, far fewer descriptions available for French, and none had the "verified" status. We decided to use this corpus nonetheless, but with the knowledge that this source for French is of a substantially lower quality (this corpus type will therefore appear as "(SCENE)" in all tables to reflect this).

e.g. *a boy is riding on a bicycle fast.* ↔ *a boy rides a bike on a dirt road.*

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T01

[2] http://www.elda.org/article125.html

[3] http://www.opensubtitles.org

	Corpus statistics 500 sentence pairs		Annotator agreements 50 sentence pairs		Tokens in paraphrase statistics not considering identity paraphrases			
					sure para.		possible para.	
	# tokens	# tokens per sent.	sure para.	possible para.	% tokens	# tokens	% tokens	# tokens
ENGLISH								
TEXT	21,473	21.0	66.1	20.4	18.6	4004	12.3	2651
SPEECH	11,049	10.5	79.1	10.9	17.5	1942	31.6	3500
SCENE	7,783	7.5	80.5	35.2	10.9	851	14.0	1094
EVENT	8,609	8.0	65.3	20.5	17.5	1506	14.5	1251
FRENCH								
TEXT	24,641	24.0	64.6	16.6	29.2	7218	6.2	1527
SPEECH	11,850	11.5	82.7	20.8	22.5	2667	16.7	1981
(SCENE)	7,012	6.5	42.8	9.3	3.9	275	9.4	664
EVENT	9,121	9.1	67.8	3.8	19.6	1793	9.6	876

Table 1: Description of all corpora and paraphrase reference sets for English (top) and French (bottom). Note that SCENE for French appears within parentheses as we do not consider it of the same quality as the other corpora.

EVENT We used titles of news article clusters from the Google News[4] news aggregation service. We further refined the clustering algorithm by filtering out article pairs whose publication dates differed from more than one day. We repeated the same selection procedure as for TEXT and SCENE to have a maximal cluster coverage and select more similar pairs first.

e.g. *Pigeons Have an Understanding of Numbers on Par With Primates* ↔ *Pigeons Have Numerical Abilities Just Like Primates*

Table 1 provides various statistics for these corpora. The first observation is that TEXT contains significantly larger sentences than the other types, more than twice as long as those of SPEECH. Annotation was performed following the guidelines proposed by Cohn *et al.* (2008)[5] using the YAWAT tool (Germann, 2008), except that alignments where not initially obtained automatically so as not to bias our annotators' work (there were two annotators per language). The main guidelines that they had to follow were that *sure* and *possible* paraphrases must be distinguished, smaller alignments were to be prefered but any-to-any alignments may be used, and sentences should be aligned as much as possible. Henceforth, we will only consider for all reported statistics and experiments those paraphrases that are not identity pairs (e.g. (*a nice day* ↔ *a nice day*)), as they are considered trivial as far as acquisition is concerned.

Table 1 also reports inter-annotator agreement[6] values computed on sets of 50 sentence pairs. We find that acceptable values are obtained for sure paraphrases, but that low values are obtained for possible paraphrases. This was somehow expected, given the many possible interpretations of possible paraphrases, but was not a problem for our experiments: as we will describe in section 3.1, the evaluation metrics we use will not count them as expected solutions, but will simply not count them as false when proposed as candidates.

Table 1 finally shows proportions and absolute numbers of paraphrases of each type for all corpora. We find that there are approximately the same total number of paraphrases for English (16,799) and French (17,001), but that English corpora collectively have an equivalent number of sure and possible paraphrases (8,303 vs. 8,496) and French have more sure paraphrases (11,953 vs. 5,048). This may be explained by the fact that our annotators worked independently and that the corpora used have differences by nature, as our experiments will show. Other salient results include the fact that TEXT contains more sure paraphrases in number than the other corpora, that SPEECH contains relatively more possible paraphrases than the other corpora, and that SCENE has significantly fewer paraphrases, both in proportion and number. In Figure 2 various mea-

[4] http://news.google.com

[5] See http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_guidelines.pdf

[6] For each paraphrase type, we used the average of recall values obtained for each annotator set as the reference .

	synonymy	typography	tense	inclusion	pragmatics	syntax	morphology	number
ENGLISH								
TEXT	51.2	7.6	5.1	12.1	0.6	4.4	12.1	6.4
SPEECH	39.8	25.6	3.5	12.3	1.7	3.5	3.5	9.7
SCENE	50.0	1.3	13.5	21.6	0.0	1.3	5.4	6.7
EVENT	36.9	15.0	8.2	19.1	1.3	6.8	6.8	5.4
FRENCH								
TEXT	46.9	9.0	8.7	2.1	3.6	6.6	3.0	19.8
SPEECH	45.5	14.2	8.0	8.0	2.6	11.6	3.5	6.2
(SCENE)	46.4	5.3	3.5	8.9	0.0	5.3	0.0	30.3
EVENT	28.3	19.7	6.1	16.0	7.4	8.6	7.4	6.1

Table 2: Percentages of paraphrase classes in 50 randomly selected sentence pairs for reference paraphrases for English (top) and French (bottom). Classes are illustrated by the following examples: (*mutual understanding* ↔ *consensus*) (**synonymy**), (*California* ↔ *CA*) (**typography**), (*letting* ↔ *having let*) (**tense**), (*Asian Development Bank* ↔ *Asian Bank*) (**inclusion**), (*police dispatcher* ↔ *woman*) (**pragmatics**), (*grief-stricken* ↔ *struck with grief*) (**syntactic**), (*Viet-name* ↔ *Vietnam*) (**morphology**), (*mortgage* ↔ *mortgages*) (**number**).

sures of sentence pair similarities are given. TEXT contains the most similar sentence pairs according to all metrics, with EVENT at a similar level on French. SCENE has sentence pairs that are more similar than those in SPEECH for English, but this is not the case for French. While the metrics used can only provide a crude account of semantic equivalence at the sentence level, these results clearly indicate that translating from text yields more similar sentences than translating from speech.

Table 2 provides a typology of paraphrases found in all our corpora and two languages, where each class has been quantified with respect to the reference alignments.[7] The main observation here is that phrasal **synonymy** (e.g. *mutual understanding* ↔ *consensus*) is the most present phenomenon. It is also interesting to note that the EVENT corpus type, which is easy to collect on a daily basis, contains reference paraphrases spread over all classes. Lastly, it is expected that paraphrases in the **pragmatics** class (e.g. *police dispatcher* ↔ *woman*) would be difficult to acquire, as this would often rely on document context and costly world knowledge.[8]

[7]Note that typologies of paraphrases have already been proposed in the literature (e.g. (Culicover, 1968; Vila et al., 2011)), but that the choice of our classes has been primarily motivated by potential subsequent uses of the acquired paraphrases (paraphrases could be annotated as belonging to more than one class). Note also that our experiments will also include results focused on the **synonymy** class only (cf. Table 5).

[8]Reusing such types of paraphrases into applications would however often be too strongly context-dependent.

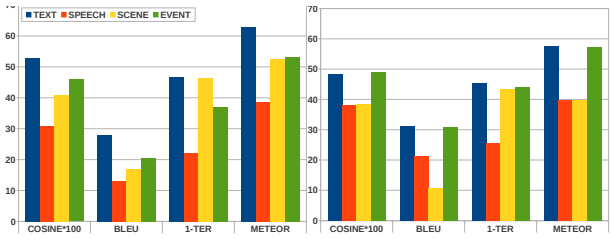

Figure 2: Sentence pair average similarities for all corpora for English (left) and French (right) using the cosine of token vectors, BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Lavie and Agarwal, 2007).

3 Bilingual experiments across corpus types

3.1 Evaluation of paraphrase acquisition

We followed the PARAMETRIC methodology described in (Callison-Burch et al., 2008) for assessing the performance of systems on the task of subsentential paraphrase acquisition. In this methodology, a set of paraphrase candidates extracted from a sentence pair is compared with a set of reference paraphrases, obtained through human annotation, by computing usual measures of *precision* (P) and *recall* (R). The first value corresponds to the proportion of paraphrase candidates, denoted $\mathcal{H}$, produced by a system and that are correct relative to the reference set containing *sure* and *possible* paraphrases, denoted $\mathcal{R}_{\text{all}}$. Recall is obtained by measuring the proportion of the reference set of *sure* paraphrases,

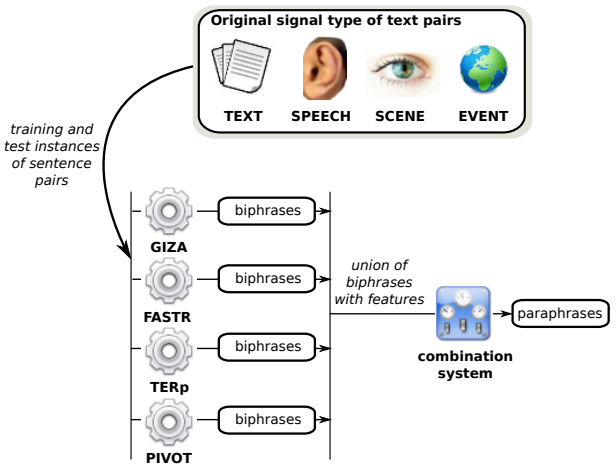

Figure 3: Architecture of our combination system for paraphrase identification.

denoted $\mathcal{R}_{\text{sure}}$, that are found by a system. We also computed an F-measure value (F_1), which considers recall and precision as equally important. These values are thus given by the following formulae:

$$P = \frac{|\mathcal{H} \cap \mathcal{R}_{\text{all}}|}{|\mathcal{H}|} \quad R = \frac{|\mathcal{H} \cap \mathcal{R}_{\text{sure}}|}{|\mathcal{R}_{\text{sure}}|} \quad F_1 = \frac{2PR}{P+R}$$

Note that the way the sets $\mathcal{R}_{\text{all}}$ and $\mathcal{R}_{\text{sure}}$ of reference paraphrase pairs are defined ensures that paraphrase pair candidates that include possible reference paraphrases will not penalize precision while not increasing recall.

All performance values reported in the following sections will be obtained using 10-fold cross-validation and averaging the results on each sub-test. All data sets of cross-validation contain 500 sentence pairs per corpus type, and 125 pairs are kept for development.

3.2 A framework for sub-sentential paraphrase identification

We now describe the systems that will be tested on the various corpora described in section 2 using the methodology described in section 3.1. Following (Bouamor et al., 2012), a combination system is used to automatically weight paraphrase pair candidates produced by individual systems using a set of features aiming at recognizing paraphrases, as illustrated on Figure 3. Four individual systems have been used and are described below: the reasons for considering those systems include their free availability, the possibility of using comparable resources when relevant for our two languages, and the specific characteristics of the techniques used.

Statistical learning of word alignments (GIZA) The GIZA++ tool (Och and Ney, 2004) computes statistical word alignment models of increasing complexity from parallel corpora. It was run on each monolingual corpus of sentence pairs in both directions, symmetrized alignments were kept and classical phrase extraction heuristics were applied (Koehn et al., 2003), without growing phrases with unaligned tokens.

Linguistic knowledge on term variation (FASTR) The FASTR tool (Jacquemin, 1999) spots term variants in large corpora, where variants are described through metarules expressing how the morphosyntactic structure of a term variant can be derived from a given term by means of regular expressions on morphosyntactic categories. Paradigmatic variation can also be expressed with constraints between words, imposing that they be of the same morphological or semantic family using existing resources available in our two languages. Variants for all phrases from one sentence of a pair are extracted from the other sentence, and the intersection of the sets for both directions is kept.

Edit rate on word sequences (TER$_p$) The TER$_p$ tool (Snover et al., 2010) can be used to compute an optimal set of word and phrase edits that can transform one sentence into another one.[9] Edit types are parameterized by one or more weights which were optimized towards F-measure by hill climbing with 100 random restarts using the held-out data set consisting of 125 sentence pairs for each corpus type.

Translational equivalence (PIVOT) We exploited the paraphrase probability defined by Bannard and Callison-Burch (2005) on bilingual parallel corpora. We used the Europarl corpus[10] of parliamentary debates in English and French, consisting of approximately 1.7 million parallel sentences, using each language as source and pivot in turn. GIZA++

[9] Note that contrarily to what TER$_p$ allows, we did not used the possibility of using word or phrase equivalents as those are only made available for English. This type of knowledge is however captured in part by the FASTR and PIVOT systems.

[10] http://statmt.org/europarl

Phrase pair features – edit distance between paraphrases, stem identity, bag-of-tokens similarity, phrase length ratio
Sentence pair features – sentence pair similarity (cosine, BLEU, TER, METEOR), relative position of paraphrases, presence of common tokens at paraphrase boundaries, presence of another paraphrase pair from each system at paraphrase boundaries, presence of a paraphrase at a different position in the other sentence
Distributional features – similarity of token context vectors for each phrase of a paraphrase (derived from counts in the large English-French parallel corpus from WMT'11 (`http://www.statmt.org/wmt11/translation-task.html`) (approx. 30 million parallel sentences)
System features – combination of the individual systems that proposed the paraphrase pair

Table 3: Features used by our classifiers. Discretized intervals based on median values are used for real values, and binarized values are used for combinations.

was used for word alignment and phrase translation probabilities were estimated from them by the MOSES system (Koehn et al., 2007). For each phrase of a sentence pair, we built its set of paraphrases, and extracted its paraphrase from the other sentence with highest probability. We repeated this process in both directions, and finally kept for each phrase its paraphrase pair from any direction with highest probability.

Automatic validation of candidate paraphrases Taking the union of all paraphrase pair candidates from all the above systems for each sentence pair, we perform a Maximum Entropy two-class classification[11], which allows us to include features that were not necessarily exploited or straightforward to exploit by individual systems to determine the probability that each candidate is a good paraphrase. More generally, this allows us to attempt to learn a more generic characterization of paraphrases, which could trivially accept any number of systems as inputs. Positive examples for the classifier are those from the union of candidates that are also in the reference set $\mathcal{R}_{\text{sure}}$, while negative examples are the remaining ones from the union. The features that we used are summarized in Table 3.

3.3 Experimental results

Results for individual systems, their union and our validation system trained on each corpus type are given on Table 4. First, we find that all individual systems fare better on TEXT, for which more training data were available and where semantic equivalence of sentence pairs is most likely. EVENT appears to be the most difficult corpus type, whereas one could say that being the most readily data source this is a disapointing result: we will return to this in section 3.4. In terms of performance on F-measure per corpus type, GIZA performs best for TEXT and SPEECH, containing long sentences with possible repetitions, while TER_p performs on par with GIZA for SCENE and best for EVENT, where equivalences that are rare at the corpus level are more present. FASTR achieves a very low recall, showing that the encoded definitions of term variants do not cover all types of paraphrases, and also possibly that the lexical resource that it uses has incomplete coverage. It nonetheless obtains high precision values, most notably on TEXT. One last comment regarding individual systems is that PIVOT is by far the most precise of all the techniques used, but with a recall much lower than those of GIZA and TER_p: as is the case for FASTR, which makes use of manually-encoded lexical resources, PIVOT encodes in some sense some kind of semantic knowledge.[12]

In all cases, our combination system manages to increase F-measure substantially over the best individual system for a corpus type and the simple union. Improvements are strong on TEXT (resp. +12.5 and +11.6 on English and French) and on SPEECH (+11.7 and +11.1) and quite good on SCENE (+3.2 and +6.4) and on EVENT (+5.4

[11] Using the implementation at: `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

[12] Note that the fact that English and French were used as the pivot for one another may have had some positive effect here, but, incidentally, the two corpora obtained by translating from the other language (TEXT and SPEECH) are not those where PIVOT fares better. The difference observed may however lie in the higher complexity of the sentences in these corpus types.

	Individual systems												Combination systems					
	GIZA			FASTR			$\text{TER}_{p\rightarrow F}$			PIVOT			union			validation		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ENGLISH																		
TEXT	48.2	58.9	53.0	63.1	5.9	10.7	41.2	66.4	50.9	**73.4**	25.8	38.2	20.8	**80.8**	33.1	68.4	62.8	**65.5**
SPEECH	39.7	44.2	41.8	27.1	3.5	6.3	25.0	50.3	33.4	**79.2**	15.3	25.7	25.5	**71.4**	37.6	51.0	56.3	**53.5**
SCENE	44.8	57.7	50.5	47.4	5.2	9.5	40.1	67.9	50.4	**84.6**	14.6	25.0	36.2	**83.4**	50.5	44.9	66.8	**53.7**
EVENT	19.0	33.9	24.3	62.9	3.1	6.0	28.8	68.7	40.6	**97.4**	11.2	20.1	20.8	**75.5**	32.7	35.0	67.1	**46.0**
FRENCH																		
TEXT	52.5	58.9	55.5	56.9	4.9	9.1	46.4	61.4	52.8	64.5	30.3	41.2	41.5	**77.9**	54.1	**74.7**	61.0	**67.1**
SPEECH	44.0	54.9	48.9	30.7	4.3	7.6	34.8	60.2	44.1	**75.5**	19.0	30.4	31.4	**76.2**	44.5	60.2	59.7	**60.0**
(SCENE)	14.4	43.6	21.7	53.0	4.0	7.4	13.8	75.3	23.4	**94.6**	5.21	9.8	12.7	**86.4**	22.2	19.9	59.8	**29.8**
EVENT	28.7	44.2	34.8	34.4	2.3	4.3	29.9	58.9	39.7	**79.5**	15.0	25.2	25.2	**72.5**	37.4	40.0	56.3	**46.8**

Table 4: Evaluation results for individual systems (left) and combination systems (right) on all corpus types for English (top) and French (bottom). Values in bold are for highest values for a given metric for each corpus type and language.

and +6.1). Recall from Table 1 that TEXT and SPEECH were the two corpus types with the highest number of sure paraphrase examples for both languages: results show that our classifier was able to efficiently use them.

Recall values for the union are quite strong for all corpus types, ranging from 71.4 (SPEECH in English) to 83.4 (SCENE in English). There is, however, a substantial decrease between the unions and the results of our combination systems, although recall values for our systems are roughly between 56 and 67, which may be considered an acceptable range on such a task. Further study of false negatives should help with engineering new features to improve paraphrase recognition. Lastly, we note that precision is in general highest for a specific system (PIVOT), and reaches high values for our validation system on TEXT, where we have the most examples (resp. 68.4 and 74.7 for English and French).

As seen in Table 2, synonymy is the most present phenomenon in all our corpora; it is also probably one of the most useful type of knowledge for many applications. We now therefore focus on this class, for which all the sure paraphrases in our corpora falling in this class have been annotated. Table 5 shows F-measure values for the individual techniques and our combination systems on all corpus types. We first observe that our combination system also always improves here over the best individual system, albeit not by a large margin on EVENT.

	GIZA	FASTR	TER_p	PIVOT	validation
ENGLISH					
TEXT	52.2	6.1	47.3	47.1	68.1
SPEECH	42.6	5.0	30.3	39.5	54.9
SCENE	51.8	6.0	48.0	26.0	56.3
EVENT	22.5	2.1	34.8	24.7	35.5
FRENCH					
TEXT	55.3	3.9	50.7	50.5	70.3
SPEECH	49.8	1.6	40.9	36.2	57.2
(SCENE)	19.6	4.2	23.1	0.0	24.7
EVENT	36.8	3.5	35.3	25.6	39.9

Table 5: F-measure values for test instances in the **synonymy** class (see Table 2) for all individual systems and our validation system for English (top) and French (bottom).

Also, we find that PIVOT performs relatively closer to GIZA and TER_p on TEXT and SPEECH than for the full set of classes, confirming the intuition that translational equivalence may be appropriate to recognize synonymy.

3.4 Experiments across corpus types

To test how different the corpora under study are as regards paraphrase identification, we now consider using as additional training data for our classifiers corpora of the other types, both individually and collectively. Results are given on Table 6.[13]

[13]Note that our results are still given by performing cross-validation averaging over 10 test sets for each tested corpus type.

	+TEXT	+SPEECH	+SCENE	+EVENT	+All
ENGLISH					
# ex+	7,342	2,296	1,784	1,171	12,593
TEXT	65.5	**66.2**	65.1	**66.2**	65.1
SPEECH	56.0	53.5	52.8	54.8	**56.6**
SCENE	49.7	**54.3**	53.7	53.8	42.7
EVENT	51.1	45.3	42.5	46.0	**56.2**
FRENCH					
# ex+	12,961	3,340	966	2,160	19,427
TEXT	**67.1**	**67.2**	66.7	67.0	66.6
SPEECH	57.6	**60.0**	56.4	59.6	57.9
(SCENE)	23.7	22.0	**29.8**	23.9	21.1
EVENT	45.2	45.6	44.3	46.8	**49.3**

Table 6: Evaluation results (F_1 scores) for all corpus types for English (top) and French (bottom) when adding training material from other corpus types (values with gray background on the diagonal are when no additional training data are used). "*#ex+*" rows indicate numbers of positive paraphrase examples for each additional corpus type.

The most notable observation is that EVENT is substantially improved by using all available additional training data for English (+10.2), and to a lesser extent for French (+2.5) . It should be noted that no individual corpus type, save TEXT, individually improves results on EVENT, and that results are yet substantially improved over the use of training data from TEXT when using all available data, revealing a collective contribution of all corpus types. The second major observation is that all other corpus types seem to be quite specific in nature, as no addition of training data from other types yields any improvement (with the exception of SPEECH on English), but they often in fact decrease performance. For instance, SCENE in English is substantially negatively impacted by the use of the numerous examples of TEXT (-4 in F-measure) and even more when using all other training data (-9). This underlines the specific nature of this corpus type: independent descriptions of the same scene in a video may be worded with much variation that mostly differ from that present in other corpus types.

Our main conclusion here is therefore that all our corpora under study are quite specific in nature, but that EVENT can benefit from all training data from the other corpus types. We can further note that the fact that TEXT is almost not impacted by additional data may also be explained by the fact that this corpus type contains more than half of the total number of examples for the two languages. Finally, there are substantially more positive paraphrase examples for French (19,427) than for English (12,593).

4 Related work

Over the years, paraphrase acquisition and generation have attracted a wealth of research works that are too many to adequatly summarize here: (Madnani and Dorr, 2010) presents a complete and up-to-date review of the main approaches. Sentential paraphrase collection has been tackled from specific resources increasing the probability of sentences being paraphrases (Dolan et al., 2004; Bernhard and Gurevych, 2008; Wubben et al., 2009), from comparable monolingual corpora (Barzilay and Elhadad, 2003; Fung and Cheung, 2004; Nelken and Shieber, 2006), and even at web scale (Pasça and Dienes, 2005; Bhagat and Ravichandran, 2008).

Various techniques have been proposed for paraphrase acquisition from related sentence pairs (Barzilay and McKeown, 2001; Pang et al., 2003) and from bilingual parallel corpora (Bannard and Callison-Burch, 2005; Kok and Brockett, 2010). The issue of corpus construction for developing and evaluating paraphrase acquisition techniques are addressed in (Cohn et al., 2008; Callison-Burch et al., 2008). To the best of our knowledge, this is the first time that a study in paraphrase acquisition is conducted on several corpus types and for 2 languages. Faruqui and Padó (2011) study the acquisition of *entailment pairs* (premise and hypothesis), with experiments in 3 languages and various domains of newspaper corpora for one language. Although their work is not directly comparable to ours, they report that robustness across domains is difficult to achieve.

Laslty, the evaluation of automatically generated paraphrases has recently received some attention (Liu et al., 2010; Chen and Dolan, 2011; Metzler et al., 2011) although it remains a difficult issue. Application-driven paraphrase generation provides indirect means of evaluating paraphrase generation (Zhao et al., 2009). For instance, the field of Statistical Machine Translation has produced works showing both the usefulness of human-produced

(Schroeder et al., 2009; Resnik et al., 2010) and automatically produced paraphrases (Madnani et al., 2008; Marton et al., 2009; Max, 2010; He et al., 2011) for improving translation performance.

5 Discussion and future work

This work has addressed the issue of sub-sentential paraphrase acquisition from text pairs. Analoguously to bilingual parallel corpora, which are still to date the most reliable resources for automatic acquisition of sub-sentential translations, monolingual parallel corpora are generally regarded as very appropriate for paraphrase acquisition. However, their low availability makes searching for *less parallel* corpora a necessity. In this study, we have attempted to identify corpora of various degrees of semantic textual similarity by considering text pairs originating from various signal types. These signal types allow various degrees of freedom as to how to formulate a text: a text is read and translated into a different language (TEXT); some speech is listened to in the context of a visual story and translated into a different language (SPEECH); some action is looked at and described (SCENE); and some event that took place is concisely reported (EVENT).

The results presented in this paper have shown how these corpora differed in various aspects. First, they contain varying quantities of paraphrases that are differently distributed into paraphrase classes. Individual acquisition techniques, based on statistical learning of word alignments (GIZA), linguistic knowledge on term variation (FASTR), edit rate on word sequence (TER_p), and translational equivalence (PIVOT), for which different performances were observed among them on the same corpus type, were shown to achieve different performances across corpus types. An efficient combination of candidate paraphrases from these individual techniques exploiting additional features to characterize paraphrases has yielded substantial increases in performance on all corpus types; however, it is interesting to note that the highest amplitude in performance across corpus types was not so much on *recall* (amplitude of 10.5 on English and 4.7 on French) than on *precision* (amplitude of 33.4 on English and 34.7[14] on French). This, some other factors aside, emphasizes the fact that the correct idenfication of paraphrases is facilitated when equivalence of semantic content is more probable. Many works have accordingly attempted to identify text units that are *as parallel as possible* from large corpora, and the task of measuring semantic textual similarity, which can find many uses, has received some attention lately (Agirre et al., 2012). However, it itself relies on some knowledge on paraphrasing.

Our avenues for future work lie in three main areas. The first one is to continue our current line of work and study the impact of additional individual acquisition techniques and better characterizations of paraphrases in context, in tandem with working on identifying parallel text pairs in large corpora. Another avenue is to start from the output of high recall techniques and to attempt to characterize the contexts of possible substitution for candidate paraphrases from large corpora as a means to acquire precise paraphrases. As the examples from Table 7 show, some classes of paraphrases, and in particular in the continuum from our **synonymy** to **pragmatics** classes, require the joint acquisition of contextual information that license substitution. Lastly, we plan to apply such knowledge in text-to-text applications.

	synonymy
TEXT	*take part in* ↔ *participate in*
	great assistance ↔ *enormous help*
SPEECH	*make a deal* ↔ *come to an agreement*
	I don't care ↔ *I don't give a damn*
SCENE	*riding a bicycle* ↔ *cycling*
	lady ↔ *woman*
EVENT	*jail escapee* ↔ *prison fugitive*
	apologizes ↔ *expresses regret*
	pragmatics
TEXT	*flew in* ↔ *arrived in*
	flood-control materials ↔ *needed supplies*
SPEECH	*face* ↔ *picture*
	want to sleep ↔ *dream about sleeping*
SCENE	*a man* ↔ *someone*
	bento ↔ *food*
EVENT	*violence* ↔ *bloodshed*
	anger ↔ *emotion*

Table 7: Examples in English for the **synonymy** and **pragmatics** classes.

[14] Not considering (SCENE) for French.

Acknowledgements

The authors would like to thank the reviewers for their comments and suggestions. This work was partly funded by ANR project Edylex (ANR-09-CORD-008).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of SemEval*, Montréal, Canada.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL*, Ann Arbor, USA.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of EMNLP*, Sapporo, Japan.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*, Toulouse, France.

Delphine Bernhard and Iryna Gurevych. 2008. Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, USA.

Rahul Bhagat and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL-HLT*, Columbus, USA.

Houda Bouamor, Aurélien Max, and Anne Vilnat. 2012. Validation of sub-sentential paraphrases acquired from parallel monolingual corpora. In *EACL*, Avignon, France.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of COLING*, Manchester, UK.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL*, Portland, USA.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4).

P. W. Culicover. 1968. Paraphrase Generation and Information Retrieval from Stored Text. *Mechanical Translation and Computational Linguistics*, 11:78–88.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*, Geneva, Switzerland.

Manaal Faruqui and Sebastian Padó. 2011. Acquiring entailment pairs across languages and domains: A data analysis. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Oxford, UK.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of COLING*, Geneva, Switzerland.

Ulrich Germann. 2008. Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-HLT, demo session*, Columbus, USA.

Wei He, Shiqi Zhao, Haifeng Wang, and Ting Liu. 2011. Enriching SMT Training Data via Paraphrasing. In *Proceedings of IJCNLP*, Chiang Mai, Thailand.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL*, College Park, USA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of NAACL-HLT*, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

Stanley Kok and Chris Brockett. 2010. Hitting the Right Paraphrases in Good Time. In *Proceedings of NAACL*, Los Angeles, USA.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. PEM: A paraphrase evaluation metric exploiting parallel texts. In *Proceedings of EMNLP*, Cambridge, USA.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of AMTA*, Waikiki, USA.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. In *Proceedings of EMNLP*, Singapore.

Aurélien Max. 2010. Example-Based Paraphrasing for Improved Phrase-Based Statistical Machine Translation. In *Proceedings of EMNLP*, Cambridge, USA.

Donald Metzler, Eduard Hovy, and Chunliang Zhang. 2011. An empirical evaluation of data-driven paraphrase generation techniques. In *Proceedings of ACL-HLT*, Portland, USA.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of EACL*, Trento, Italy.

Franz Josef Och and Herman Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignement of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of NAACL-HLT*, Edmonton, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, Philadelphia, USA.

Marius Pasça and Peter Dienes. 2005. Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web. In *Proceedings of IJCNLP*, Jeju Island, South Korea.

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving translation via targeted paraphrasing. In *Proceedings of EMNLP*, Cambridge, USA.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word Lattices for Multi-Source Translation. In *Proceedings of EACL*, Athens, Greece.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, Boston, USA.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2010. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

Jörg Tiedemann. 2007. Building a Multilingual Parallel Subtitle Corpus. In *Proceedings of the Conference on Computational Linguistics in the Netherlands*, Leuven, Belgium.

Marta Vila, M. Antònia Martí, and Horacio Rodríguez. 2011. Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach. *Procesamiento del Lenguaje Natural*, (462-3).

Sander Wubben, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. 2009. Clustering and maching headlines for automatic paraphrase acquisition. In *Proceedings of the European Workshop on Natural Language Generation*, Athens, Greece.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven Statistical Paraphrase Generation. In *Proceedings of ACL-IJCNLP*, Singapore.