A Results COCO

A.1 ALL and KEEP

As for Flickr8k, the results obtained in the ALL and the KEEP condition show that the ALL condition brings little improvement over the baseline results. However, we may observe a different pattern in the RANDOM-KEEP condition: whereas for Flickr8k the results are significatively worse in this condition, some results are significatively better than the baseline for COCO (see for example the results obtained at $\text{GRU}_{\text{PACK.}}$–3,4 with phones, syllables-connected and syllables word). We explain this by the fact that, contrary to Flickr8k that used real human speech, COCO uses synthetic speech with only one voice and hence, has very low intra-speaker variation. Thus, even though we randomly subsample the input, as there is very little intra-speaker variation, the network is much more likely to figure out from which units the subsampled vector came from. Thus, randomly subsampling the spoken input acts as a form of regularisation for the network such as dropout.

A.2 Phones, Syllables, or Words

We also observe that the best results are obtained when we use word segments ($\text{GRU}_{\text{PACK.}}$–2). As for Flickr8k, word units yield more consistent results over most of the layers ($\text{GRU}_{\text{PACK.}}$–1,2,3,4) suggesting that word-like segmentation is the adequate segmentation to be used for our task. We also notice that syllables word, that preserve word boundaries, obtain results close to that of word segments. As for Flickr8k, syllables-connected overall yield worse results that phones or syllables-word, once again showing that preserving word boundaries seems to be important.

A.3 $\text{GRU}_{\text{PACK.}}$ Layer Position

Results on the COCO data set also show that the worse results are obtained when boundary information is provided at the last layer ($\text{GRU}_{\text{PACK.}}$–5) showing that this layer is not concerned with form anymore, but with semantics. The best results (be it with phones, syllables-connected, syllables-word, or word) are obtained at the second layer. Results then decrease in the upper layer.

B Hierarchical Segmentation

The results obtained with hierarchical models that use phones and syllables-word, and syllables-word and words are shown respectively in Table 6 and Table 7. Results are worse than when using either a hierarchical model with phone and word segments or a model with phone, syllable-word, and word segments. This shows that preserving low-level segments such as phones and high-level segments such as words enables the model to better generalise. Also, according to the results presented in Table 3, it appears that the architecture of the network should be deeper (5 layers) when using both phone and word segments than when using other type of segments as the best models of Table 6 and 7 converge better with 4 layers. This suggests that using phone and word segments requires an additional amount of processing in order to be used effectively. Finally, the difference (-0.3pp) in the results obtained with a phone and word architecture, and a phone and syllable-word architecture show that even though syllables-word and words are quite close in length (see compression rates in section 4.4), they are not equivalent in terms of semantic content, otherwise we would have observed identical results.

	COCO — KEEP condition								COCO — ALL condition							
GRU	Phones		Syl.-Co.		Syl.-Word		Word		Phones		Syl.-Co.		Syl.-Word		Word	
Pack.	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R
5	9.4	**9.6**	**9.1**	**9.1**	**9.6**	9.1	***9.4***	*8.7*	**9.7** $^{+}$	9.4	9.1	**9.5**	**9.5**	9.4	9.3	**9.5**
4	10.0 $^{+}$	**10.5** $^{+}$	**10.2** $^{+}$	9.6	**10.4** $^{+}$	9.9 $^{+}$	***10.6*** $^{+}$	*9.5*	**9.3**	9.2	**9.4**	9.1	**9.6**	9.2	9.0	**9.4**
3	**10.5** $^{+}$	10.1 $^{+}$	**10.4** $^{+}$	9.8 $^{+}$	**10.5** $^{+}$	10.1 $^{+}$	***11.0*** $^{+}$	*9.7*	**9.5**	9.2	**9.2**	9.1	**9.4**	9.1	**9.4**	9.2
2	***10.7*** $^{+}$	*9.8* $^{+}$	***10.5*** $^{+}$	*9.4*	***10.9*** $^{+}$	*9.3*	***11.3*** $^{+}$	*8.8*	**9.4**	8.9	**9.7** $^{+}$	9.1	**9.6**	9.2	***9.7*** $^{+}$	*8.9*
1	***10.1*** $^{+}$	*7.9* $^{-}$	***9.7***	*7.1* $^{-}$	***10.2*** $^{+}$	*7.0* $^{-}$	***10.3*** $^{+}$	*7.0* $^{-}$	**9.8** $^{+}$	9.4	**9.6**	9.4	***10.0*** $^{+}$	*9.1*	**9.5**	9.1

Table 5: Maximum R@1 (in %) for each model trained on the COCO data set. The same naming conventions of Table 2 are used for this table.

Architecture	5 layers					4 layers				3 layers			2 layers	
2^{nd}GRU$_{\text{PACK.}}$ / 1^{st}GRU$_{\text{PACK.}}$	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**1**	**2**
1		6.6	6.0	6.3	4.3		6.9	6.5	4.6		**6.6**	4.9		**4.6**
2			**7.5**	6.8	5.7			**7.9**	4.7			6.1		
3				6.5	4.8				4.7					
4					4.6									
5														
Baseline	4.3					4.4				3.4			3.5	

Table 6: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of two GRU$_{\text{PACK.}}$ layers using phones and syllable-word (models were selected based on the maximum R@1 on the validation set). The same naming conventions of Table 3 are used for this table

Architecture	5 layers					4 layers				3 layers			2 layers	
2^{nd}GRU$_{\text{PACK.}}$ / 1^{st}GRU$_{\text{PACK.}}$	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**1**	**2**
1		5.7	5.5	5.7	4.5		5.7	6.8	5.2		6.0	5.2		5.3
2			**7.3**	7.1	6.1			**7.6**	6.0			**6.3**		
3				6.8	5.7				6.0					
4					5.5									
5														
Baseline	4.3					4.4				3.4			3.5	

Table 7: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of two GRU$_{\text{PACK.}}$ layers using syllable-word and word segments (models were selected based on the maximum R@1 on the validation set). The same naming conventions of Table 3 are used for this table.