A avaliação final foi feita pela comparação dos algoritmos usando as medidas *Pk* e *WindowDiff*. É apresentada também, para fins de comparação, as medidas tradicionais acurácia, precisão, revocação e F[1], entretanto, nesse contexto, essas medidas são menos significativa que Pk e WindowDiff, conforme já mencionado na Seção **??**. A Tabela 1 contém as médias com cada algoritmo. Vale lembrar que P_k e *WindowDiff* são medidas de dissimilaridade, ou seja, os valores menores significam melhores resultados.

Método	**Pk**	**WD**	**A**	**P**	**R**	**F1**	**Segmentos**
Sentenças	0.320	0.502	0.498	0.498	**1.000**	**0.642**	22.083
TextTiling	0.275	0.469	0.531	0.514	0.937	0.640	19.583
C99	0.142	0.426	0.574	0.601	0.473	0.506	8.167
BayesSeg	0.148	0.414	0.586	0.599	0.526	0.528	8.750
MinCut	0.226	0.532	0.468	0.464	0.438	0.432	10.333
TextSeg	**0.085**	**0.387**	**0.613**	**0.714**	0.412	0.497	5.167

Table 1: Melhores resultados obtidos.

Na Figura 1 é apresentada a performance dos algoritmos nas medidas tradicionais. Observa-se valores altos de revocação para a segmentação por sentenças, pois é atribuído um limite a todo candidato a final de segmento, o que resulta no valor máximo para revocação. De maneira semelhante, o comportamento do *TextTiling* gera mais segmentos em relação aos demais, e com isso tem-se valores maiores de revocação, o que pode ser contornado configurando o algoritmo com passos maiores, ou ainda, sobre-escrevendo a função que calcula os *depth scores* para reconhecer vales mais largos.

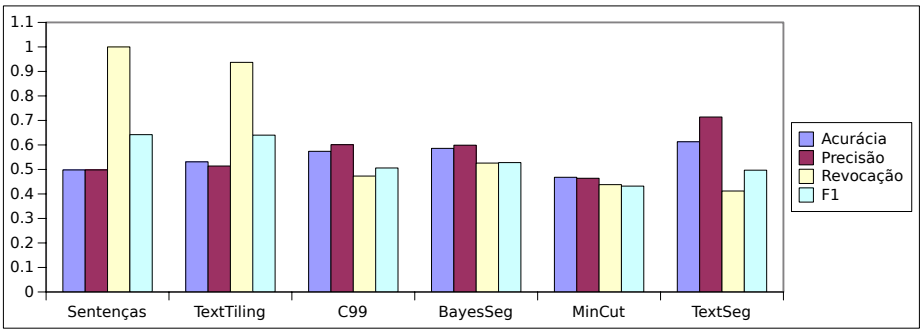

Figure 1: ...

Na Figura 2 é apresentada a performance dos algoritmos nas medidas P_k e *WindowDiff*. Verifica-se que *TextSeg* apresenta valores de *WindowDiff* próximas ao *C99* e *BayesSeg* e resultados mais significantes quando medidos por P_k em relação aos demais algoritmos.

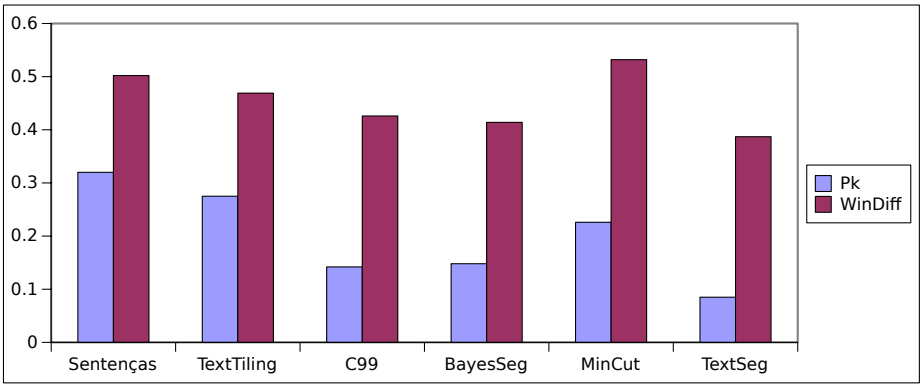
0.6
0.5
0.4
0.3
0.2
0.1
0
Sentenças
TextTiling
C99
BayesSeg
MinCut
TextSeg
Pk
WinDiff

Figure 2: ...