Corpus المدونات المستخدمة

تم استخدام المدونات الخاصة بالأمم المتحدة The United Nations Parallel Corpus v1.0 في عملية تدريب و اختبار الشبكة العصبية و تقييمها

وعي عبارة عن السجلات الرسمية, و الوثائق و المناقشات لمئسسات الأمم المتحدة و هذه السجلات مكتوبة باللغات الرسمية الست لأمم المتحدة، هذه الوثائق و السجلات قد تم جمعها و ترجمتها بشريا بين عامي 2014-1990 ، كما تم مقابلة هذه الجمل يدوياً أيضاً

تتألف مدونات الأمم المتحدة من أجل ثنائيات الجمل العربية و الإنكليزية من عدة وثائق بلغ عدد هذه الوثائق 111241 وثيقة , بلغ عدد ثنائية الجمل المتناظرة من 18,539,207، و كان عدد الكلمات العربية ضمن هذه الوثائق 456,552,223 كلمة، في حين عدد الكلمات الإنكليزية 512,087,009 كلمة كما بلغ حجم هذه المدونة 6GigaByte.

السبب الرئيسي للاعتماد على هذه المدونة