Even though the Bantu language of Swahili is spo- ken by more than fifty million people in East and Central Africa, it is surprisingly resource-scarce from a language technological point of view, an unfortunate situation that holds for most, if not all languages on the continent. The increasing amount of digitally available, vernacular data has prompted researchers to investigate the applicability of corpusbased approaches to African language technology . In this vein, the SAWA corpus project attempts to collect and de- ploy a parallel corpus English - Swahili, not only for the straightforward purpose of developing a machine translation system, but also to investigate the possibil ity of projection of annotation into a resource-scarce, African language. Compiling a balanced and expansive parallel corpus English - Swahili is a rather daunting task. While monolingual Swahili data is abundantly available on the Internet, sourcing parallel texts is cumbersome. Even countries that have both English and Swahili as their official languages, such as Tanzania, Kenya and Uganda, do not tend to translate and/or publish all government documents bilingually. One therefore op- portunistically collects whatever can be found in the public domain. At this point in the data collection phase, that means that the 2.2 million word parallel corpus is biased towards religious material, such as bible and quran translations. Nevertheless, the more interesting, secular part of the SAWA corpus (420k words) is steadily increasing, thanks to the inclusion of bilingual investment reports, manually translated movie subtitles, political documents and material kindly donated by local translators to the SAWA project. Each text in the SAWA corpus is automatically part-of- speech tagged and lemmatized, using the TreeTagger for the English part (Schmid, 1994) and the systems described in De Pauw et al. (2006) and De Pauw and de Schryver (2008) for Swahili. These extra annotation layers allow us to perform more accurate automatic word alignment on the basis of factored data