The Multisix Corpus

While in the monolingual tasks at CLEF 2003 QA we had three different document collections and three sets of questions, all the bilingual tasks had one single English target corpus. We generated 200 English queries, processed them and then translated them into Dutch, French, German, Italian and Spanish. Due to time constraints we could not compile a large pool of general questions independently from the corpus and then verify them. Instead, we chose an alternative approach: we randomly selected a document from the collection (trying to select news with a general importance, avoiding sections that deal with local politics or issues too strictly related to Los Angeles counties) and picked up a text snippet that was relevant, long and interesting enough to get a question out of it. For instance, from the following passage

Ayrton Senna, a three-time world Formula One driving champion whom some racing observers called the best race driver of all-time, died Sunday of head injuries suffered when his car crashed during the Grand Prix of San Marino in Imola, Italy.

we drew the following question:

<qa cnt="43" type="LOCATION">
  <language val="ENG" original="TRUE">
    <question assessor="Ale-irst">
      Where did Ayrton Senna have the accident that caused his death?
    </question>
    <answer n="1" idx="LA050294.0187">
      in Imola, Italy
    </answer>
  </language>
  <language val="DUT" original="FALSE">
    <question assessor="">
      Waar verongelukte Ayrton Senna?
    </question>
    <answer n="1" idx="">
      SEARCH[in Imola, Italy]
    </answer>
  </language>
  <language val="FRE" original="FALSE">
    <question assessor="">
      Où Ayrton Senna a-t-il eu l'accident qui a entraîné sa mort?
    </question>
    <answer n="1" idx="">
      SEARCH[in Imola, Italy]
    </answer>
  </language>
  <language val="GER" original="FALSE">
    <question assessor="">
      Wo hatte Ayrton Senna seinen tödlichen Unfall?
    </question>
    <answer n="1" idx="">
      SEARCH[in Imola, Italy]
    </answer>
  </language>
  <language val="ITA" original="FALSE">
    <question assessor="">
      Ayrton Senna dove fece l'incidente che causò la sua morte?
    </question>
    <answer n="1" idx="">
      SEARCH[in Imola, Italy]
    </answer>
  </language>
  <language val="SPA" original="FALSE">
    <question assessor="">
      ¿Dónde tuvo Ayrton Senna el accidente que causó su muerte?
    </question>
    <answer n="1" idx="">
      SEARCH[in Imola, Italy]
    </answer>
  </language>
</qa>
The format we used to structure this test set is the same of the DISEQuA Corpus.

As can be noticed, the attribute `assessor` in the non-English versions of the query is empty, in fact the questions were processed and verified just in the *Los Angeles Times* collection. Similarly, the Boolean value of the attribute `original` is `TRUE` just in the English version of the queries.

The corpus of questions and answers that we obtained is rather balanced in terms of question types: it is composed of 42 queries that ask for a LOCATION, 39 for the name or role of a PERSON, 34 for a MEASURE or amount of something, 30 for a DATE, 28 for an ORGANIZATION, 22 for OTHER instances (that could be labelled with many other categories) and 5 for concrete OBJECTS. The test set includes 15 NIL questions, that have no answer in the *Los Angeles Times* document collection.