The conclusions reached in the World Cafe rounds can be summarized as follows:

Table 1: Metadata

A minimum of metadata should be provided for every corpus. Metadata should cover learner characteristics and corpus composition and annotation. *Learner* characteristics should preferably include the following:

1. L1 and knowledge of other languages,
2. country of residence,
3. gender (m, f, x),
4. age or age group,
5. education,
6. proficiency level of person and performance,
7. exposure to target language.

Corpus information should include the following:

1. mode,
2. target language,
3. size,
4. license,
5. availability,
6. task or prompt,
7. annotation (including annotators).

Furthermore, metadata should include provenance and availability (including license), as well as a persistent ID for citation.
The collection of the metadata should be standardized (maybe through automated harvesting), but no conclusions were made at that point.

Table 2: Happy user (interface, annotation)

Needs analysis is of utter importance since learner corpora are of interest to a wide range of users with varying backgrounds and varying technical expertise. The main types of users are likely to be teachers, course books designers, assessment developers, researchers (in second language acquisition, language teaching, lexicography, computational linguistics, etc), software developers and the general public. Surveys of what future users need and how they can use the resources have been performed e.g. within the MERLIN project where it was later followed up by a test-run of the platform by the different user groups so that the system could be made more user-friendly.

There is agreement on the need for the corpora to be easily accessible and exploitable in a variety ways. Web browsers are often the simplest way of accessing the resources; web interfaces therefore need to attract all types of users and support easy navigation.

Different users may need different ways of using a corpus. The new interface for ASK, demonstrated by Paul Meurers (the main site is http://clarino.uib.no/ask/ask, but Meurers also showed a new interface), has the advantage of two interacting modes. On the one hand, a simple search makes it possible to start a query without knowing details about the query

language and the query can be extended with menus. On the other hand, the simple search is also translated to a query language which can be further edited by more advanced users, if desired. A crucial point is that it is easy to switch between the different search modes.

It would be good if the simplest search option always offered possibilities to search for one or two words in a way similar to a simple Google search, since most people in the general public are used to that and therefore that should be the easiest way to start. We also think it would be good to be able to access frequencies and to be able to make your own subcorpus, e.g. if you are only interested in a group of learners who have a specific L1 or a certain age group or level.

Web interfaces should also make it clear if it is possible to download the whole corpus (which computational linguists often want to do) and if so how this can be done and what kind of licence there is. Licences should generally be non-commercial and if a corpus offers commercial licences that should usually be something that one can possibly acquired after negotiations. However this means that it is important to have a contact address for inquiries and that someone must be willing and able to take care of such inquiries on a regular basis.

Researchers usually have a need to be able to access information about how the data in the corpus have been collected in general, so this kind of information should also be easily accessible, as well as information about how it has been annotated. We agreed that a manual must be easily accessible, in addition to which video clips for different types of users which show what you can do with the corpus (cf. youtube clips for Antconc) would be very good. It would also be a good idea to offer workshops for different user groups.

We also discussed the possibility of creating exercises on the basis of corpora, e.g. of the kind offered in Lärka (www.spraakbanken.gu.se/larkalabb). We liked the look of what Lärka can offer at the moment and agreed that it would be of interest to find out how easy it would be to use the Lärkasystem on other corpora.

Corpora could also be used as a basis for spell checkers, grammar checkers, potentionally also to check the level of difficulty of a text to be read or proficiency in a text. If corpora are easily comparable it would also facilitate studies of interference in L2 from previously learned / acquired languages (L1 / L2 / FL).

We explored the kind of additional annotation that there might be an interest in and we all liked the possibility of adding your own annotation as a researcher and making that searchable, as is possible in the ASK-interface. Additional annotation that might be of interest to some may cover senses, semantic roles, constructions (CxG), information structure, complexity etc. Furthermore, metadata such as genres, registers, topic, may provide useful information.

Finally, it would be of interest to have a possibility to collect ideas from users about what they might need, but also of ideas they have had about how the corpus can be used (e.g. a possibility to collect a database of articles based on corpus material, exercises based on

corpora, teacher's notes on how corpora can be used in teaching etc. cf the Language Change Database http://www.helsinki.fi/lcd/).

Table 3: Error-annotation

The most important part of the error-annotation should be the needs analysis, to determine who are the end users of the error-annotated learner corpus, so that the error-annotation can be driven by research questions.

We should aim to formulate some general principles (guidelines) with respect to the goals of the error-annotation in order to achieve the maximum usability for a variety of research questions. The annotators need to receive training on the guidelines, with as many examples as possible.

In other to establish interoperability, we should all agree on some kind of major types/categories of errors. After the brief analysis, we discovered that we all have some general shallow-level categories in our error-annotation taxonomies (e.g. spelling, grammar, lexis and semantics), but a cross-language comparison would be beneficial, as well as the adaption of the shallow-level taxonomy to different languages.

All tag sets that are developed should firstly be used by developers, and after that the logic behind tags needs to be analyzed, since some tags might be misleading for the annotator. The need for a huge tag set could be limited by linguistic annotations that are present in almost all learner corpora.

The annotators should avoid to annotate the cause of the error (such as the L1 influence), but should rather concentrate on the description of the error itself. Since error-annotation tends to be expensive and time-consuming process, one can do error correction and annotation separately, in a way that native speakers (who are not linguist specialists) perform correction, while linguist specialist (both native and non-native speakers) perform the annotation.

Finally, it is also important to establish the profile of an annotator (especially for the purpose of the inter-annotator agreement).

Table 4: Tools and software

The general conclusion regarding tools for L2 corpora was that there is no clear lack of specific new tools, but there is a general desire to have less computer savvy interfaces to existing tools. Moreover, there are currently three main obstacles with tools: firstly, it is difficult to know which tools are out there, which especially for people just starting in the field makes it hard to find tools that would suit them. The second issue is that too many of the tools are not available for those who would want to use them. And the third issue is that many of the tools out there are not properly documented.

To address the first issue, it was decided to produce not only a list of existing L2 corpora after this workshop, but also work on a list of tools, with a description of what they do, whether they are available, and whether they are still maintained. All participants will be asked to complete that list after it has been set up.

The second issue cannot be easily resolved. There are typically two reasons why a tool is not available: either because they were specifically designed for a single project and basically restricted to that project, or if the tool is designed so that it can be used outside of the scope of the project it was designed for, the author has not made it available, either because he does not have the permissions to do so, or because doing so entails a lot work (answering questions, providing an installer, creating documentation, debugging for different servers, etc.).

Even those tools that have been made available too often get discontinued after the author has moved on to other things, meaning they will stop working over time. Resolving this would involve looking for a financing structure in which key tools are financed by the community that uses it rather than by a single project, or have key tools under the roof of a commercial firm, as for instance in the case of SketchEngine, which will have a commercial incentive to keep the tool alive and updated.