Verteilte Korpusabfragesysteme

Tobias Roth (Basel)

Abstract

Distributed text corpora have not been very much in use so far. The Swiss Text Corpus (CHTK) and its partner projects set up a distributed corpus for German ("Korpus C4"), virtually merging parts of their corpus data and making them available through one common query platform.

Based on experience made during this project, we propose a possible path towards a more standardised interface for distributed corpus queries. This should allow to integrate new as well as existing corpora more easily into distributed corpus systems. Special attention is paid to problems such as responsibility assignment, performance, user management, format unification and metadata synchronisation.

1 Einleitung

Textkorpora wie auch andere linguistische Ressourcen sind oft örtlich stark verteilt und selbst in der heutigen Zeit, wo dies technisch machbar wäre, nicht immer problemlos und vollumfänglich öffentlich zugänglich. In vielen Fällen nutzt man mehrere Korpora neben-einander – sei es um eine weitere Vergleichsbasis oder schlicht um mehr Daten zur Verfügung zu haben. Eine gemeinsame Nutzung mehrerer Korpora gleichzeitig wäre sehr erwünscht. Analog zum Konzept der Metasuchmaschine im Web wären verteilte Korpora einfacher und besser nutzbar, wenn sie verteilt abgefragt werden könnten, am besten über standardisierte Schnittstellen.

In diesem Artikel soll in einem ersten Teil kurz der Bedarf für solche Schnittstellen aufgezeigt werden (cf. Abschnitt 2). Anschliessend wird das Projekt Korpus C4 vorgestellt, das in den letzten Jahren ein deutschsprachiges, verteiltes Korpus aufgebaut hat und an dem das Schweizer Textkorpus beteiligt ist. Es sollen dabei vor allem Eigenschaften, Erfahrungen und Probleme mit besonderem Bezug zur verteilten Abfrage dargestellt werden (cf. Abschnitt 3).

Daraufhin wird skizziert, wie ausgehend von den Erfahrungen mit dem Korpus C4 eine stärker standardisierte Schnittstelle aufgebaut sein könnte und worauf dabei besonders zu achten wäre (cf. Abschnitt 4).

2 Motivation für verteilte Korpusabfragesysteme

Es besteht eine Vielzahl unterschiedlicher Korpusprojekte, die sich in vielen Punkten aber doch sehr ähnlich sind. Der Zugang zu den Daten geht heute oft über eine webbasierte Abfrageoberfläche. Frei und komplett zugänglich sind die Korpusdaten auf diese Weise in den meisten Fällen nicht. Urheberrechtliche, datenschutztechnische oder projektstrategische Gründe stehen dem entgegen: Korpusprojekte dürfen und wollen ihre Daten nicht einfach ungeschützt im Volltext preisgeben.

Wer in seiner Forschung mehrere Korpora gleichzeitig nutzen möchte, kommt deshalb nicht darum herum, sich mit den Abfrageoberflächen der einzelnen Projekte auseinanderzusetzen. Mit jeweils leicht unterschiedlicher Abfragesprache, Benutzerführung und Präsentation der

Resultate wird sehr ähnliche Funktionalität angeboten. Für eine Abfrage an mehrere Korpora sind Forschende gezwungen, diese in jeweils unterschiedlicher Syntax an die verschiedenen Korpusportale abzusetzen, die einzelnen Resultate zu sammeln, zu interpretieren und bei entsprechender Vergleichbarkeit zusammenzuführen. Gewisse Arten von Abfragen wie beispielsweise die Suche nach Kollokationen sind zudem sehr schlecht als Einzelabfragen an verschieden Korpora mit nachträglicher Zusammenführung möglich, da sie als Vergleichsbasis die Gesamtheit der Daten benötigen.

Es ist offensichtlich, dass ein vermehrtes Zusammenführen von Korpora zu grösseren Korpora und der Zugang dazu über eine einzige Abfrageoberfläche und -syntax für viele korpuslinguistische Fragestellungen eine Vereinfachung und eine erleichterte Verfügbarkeit relevanter Daten bedeuten würde. Das *Korpus C4*, eine gemeinsame Initiative des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS), des Austrian Academy Corpus (AAC), des Korpus Südtirol und des Schweizer Textkorpus (cf. Abschnitt 3), hat sich unter anderem aus solchen Gründen zum Ziel gesetzt, ihre Korpora bzw. Teile davon, zu einem gemeinsamen Korpus zusammenzuschliessen.

Ein grosses Hindernis beim Zusammenschluss von Korpora verschiedener Institutionen ist jedoch das oben angetönte und durchaus verständliche Festhalten an der Datenhoheit. Diese verhindert für einen Zusammenschluss mehrerer Korpora die technisch einfachste Lösung, nämlich das Kopieren der Daten an einen einzigen Standort, wo sie ganz klassisch als ein Korpus für die Webabfrage aufbereitet werden könnten.

Um dieses Problem zu umgehen, bietet es sich an, Korpora virtuell über verteilte Abfragen der einzelnen Teilkorpora zusammenzuschliessen. Die Abfragen werden dabei von einer zentralen Abfrageoberfläche aus an die beteiligten Korpora weitergeleitet. Diese senden ihre Resultate zurück, die von der Abfrageoberfläche vereint und einheitlich präsentiert werden. Der Vorteil für die Nutzerschaft liegt darin, dass über eine Schnittstelle mehrere Korpora gleichzeitig abgefragt werden können. Die Korpusprojekte ihrerseits müssen dazu nicht ihre Daten unkontrolliert aus der Hand geben und den beteiligten Partnerkorpora kaum mehr Rechte vergeben als ihren eigenen Nutzerinnen und Nutzern.

Neben bestehenden Formen der Kooperation und Standardisierung in der Korpuslinguistik – man denke z.B. an das Kindersprachenkorpus *CHILDES* (cf. MacWhinney 1991), das Text-Engineering-Framework *GATE* (cf. Cunningham et al. 2002), das Korpusportal des Wortschatzprojekts in Leipzig (cf. Quasthoff/Richter/Biemann 2006) oder das Grid-Projekt *TextGrid* (cf. Neuroth/Kerzel/Gentzsch 2007) – können verteilte Korpusabfragesysteme eine sinnvolle Ergänzung sein. In eine ähnliche Richtung hin zur Vernetzung von linguistischen Ressourcen gehen auch die momentan laufenden Projekte CLARIN (cf. http://www.clarin.eu) und D-SPIN (cf. http://www.sfs.uni-tuebingen.de/dspin).

3 Korpus C4

Das verteilte Korpus, das mit Beteiligung des Schweizer Textkorpus in den letzten Jahren aufgebaut wurde, konnte kürzlich unter dem Namen *Korpus C4* (cf. http://www.korpus-c4.org) aufgeschaltet werden. Daran beteiligt sind die Korpora des DWDS aus Deutschland ('Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts' – http://www.dwds.de), des AAC aus Österreich ('Austrian Academy Corpus' – http://www.aac.ac.at), des Korpus Südtirol (http://www.korpus-suedtirol.it) und des Schweizer Textkorpus (http://www.schweizer-textkorpus.ch, cf. auch Bickel et al. 2009).

Ziel war ein gemeinsames Textkorpus der vier Partnerinstitutionen. Recht schnell wurde bei der konkreten Umsetzung klar, dass es sehr schwierig werden würde, die Daten physisch zusammenzulegen, genau aus den in Abschnitt 2 genannten rechtlichen Gründen. Die Partner

einigten sich deshalb auf ein verteilt abfragbares Korpus. Die Idee eines verteilten Korpus stand also nicht am Anfang, war eher aus der Not heraus geboren. Konsequenterweise wurde für deren Realisierung ein sehr pragmatischer Ansatz verfolgt: Oberste Priorität hatte ein funktionierendes, gemeinsames Korpus und nicht der Nachweis der Machbarkeit eines verteilten Korpus. Es wurde deshalb immer viel weniger Energie auf eine möglichst allgemeine Schnittstelle als auf das konkrete Funktionieren des Korpus verwendet. Die Realisierung erfolgte nach einem Bottom-up-Ansatz. Es wurde zuerst die einfachste Form einer verteilten Abfrage unter nur einem Korpusserver realisiert. Diskussionen um allgemeinere Schnittstellen waren auch geführt worden. Ihre Umsetzung wurde aber hintangestellt und ist in dieser Projektphase auch nicht mehr begonnen worden.

Doch obwohl das Korpus C4 nur auf das Zusammenspiel der beteiligten vier Korpora fokussiert war und explizit keine allgemeine Standardabfrageschnittstelle definierte, wie sie für die Integration weiterer Korpora oder für den Zusammenschluss beliebiger Korpora angezeigt wäre, können die Erfahrungen aus dem Bau des Korpus C4 wichtige Hinweise liefern, wo auch bei einer allgemeineren Standardabfrageschnittstelle Probleme und Hindernisse auftreten könnten. Nachfolgend werden wichtig Charakteristika des Korpus-C4-Abfragesystems kurz vorgestellt sowie Probleme und Hindernisse bei seinem Aufbau angesprochen.

3.1 Einheitliches System

Damit eine verteilte Abfrage möglich wird und die zurückgelieferten Resultate einheitlich dargestellt werden können, müssen die Teilkorpora einheitlich angesprochen werden können. Da im Rahmen von C4 wie gesagt keine allgemeine Standardabfrageschnittstelle für Korpora definiert wurde, mussten die Teilkorpora in den relevanten Punkten möglichst viel vereinheitlichen – das heisst, wo immer möglich, dieselbe Software und dasselbe Datenformat verwenden. Kernstück hier ist die Verwendung einer einzigen Indexierlösung/Suchmaschine. Es wurden zwar im Projektverlauf auch Varianten mit gemischten Indexierlösungen diskutiert, doch wurde dann beschlossen, das Korpus in einem ersten Schritt unter einer einzigen Indexierlösung laufen zu lassen. Die Wahl fiel dabei auf die linguistische Open-Source-Suchmaschine *DWDS/Dialing Concordance (DDC)*, entwickelt von unserem Partnerprojekt in Berlin (cf. Sokirko 2005).

Hauptgründe für diese Wahl waren einerseits der mehrjährige stabile Einsatz unter Produktivbedingungen bei den Korpora des DWDS, andererseits und wohl ausschlaggebend aber die Tatsache, dass DDC die Möglichkeit des Einsatzes verteilter Korpora schon von Haus aus bietet. Abbildung 1 zeigt eine grobe Übersicht der daraus resultierenden Architektur des Gesamtsystems sowie den Weg einer Abfrage. Die Knoten A-D repräsentieren die Korpusserver der einzelnen Forschungsstellen, die alle gleich aufgebaut sind. Eine Anfrage eines Benutzers an Knoten A geht über die Abfrageoberfläche von Knoten A (C4 WebApp) zum eigenen DDC-Server, der die Abfrage an alle beteiligten Teilkorpora (DDC-Thread) auf allen Knoten weiterleitet und die Resultate zusammengeführt an die Abfrageoberfläche zurückreicht, die sie schliesslich darstellt. Auf der Abbildung wird deutlich, dass insbesondere die Kompatibilität der vier DDC-Threads untereinander unabdingbar ist, und Kompatibilität bedeutet in diese Fall, dass sie in Softwareversion und Datenformat (cf. Abschnitte 3.1.1 und 3.1.2) identisch sind.

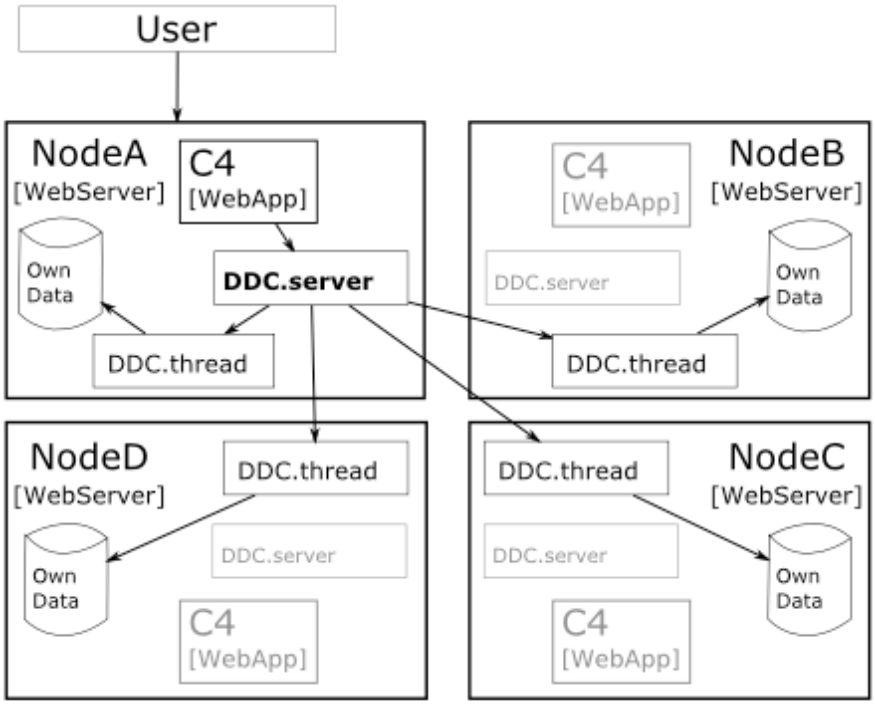

Abb. 1: Architekturübersicht Korpus C4, DDC-basiert (Diagramm: Matej Durco)

3.1.1 Einheitliche Metadaten

Ebenfalls vereinheitlicht innerhalb der Teilkorpora des Korpus C4 wurden die indexierten und damit durchsuchbaren Metadaten. Der gewählte Ansatz vereinheitlichter Systeme machte diesen Schritt unabdingbar. Die Einigung auf eine gemeinsame Metadatendefinition gestaltete sich nicht ganz einfach, sie konnte aber schliesslich doch in einer gemeinsamen DDC-Options-Datei (cf. Sokirko 2005: 4ff.) konkret festgehalten werden. Die Einigung bedeutete, nicht weiter verwunderlich, in den meisten Fällen eine Reduktion auf den kleinsten gemeinsamen Nenner unter den vier Projekten. Im gemeinsamen Korpus indexiert werden Metadaten wie Publikationsjahr, Autor, Titel, Quellennachweis, Publikationsregion, Werkkategorie[1] und Textsorte.

Man kann sich leicht vorstellen, dass bei zusätzlichen beteiligten Projekten eine Einigung noch schwieriger werden und die Zahl der gemeinsam indexierbaren Felder weiter schrumpfen könnte.

3.1.2 Einheitliche Formate

Mit der Vereinheitlichung der Formate sind vor allem zwei Bereiche angesprochen. Einerseits muss das Textformat dem entsprechen, was die Indexierlösung DDC mit der definierten Options-Datei als Eingabeformat erwartet, d. h. im Falle des Korpus C4 eine XML-Datei mit TEI-konformem[2] Header mit den definierten Metadaten (cf. 3.1.1) sowie dem Textteil mit einem Texttoken pro Zeile inkl. den Metadaten des jeweiligen Tokens. Folgender Auszug

[1] Belletristik, Sachtexte, Gebrauchstexte oder journalistische Prosa.

[2] TEI steht für *Text Encoding Initiative* – ein Konsortium, das Standards für die Kodierung digitaler Texte in XML definiert. Die vom TEI-Konsortium herausgegebenen Guidelines geben vor, welche Auszeichnungselemente mit welcher Bedeutung verwendet werden sollen, um eine optimale Kompatibilität zu erreichen – Kompatibilität von Texten untereinander für leichteren Austausch, aber auch von Texten zu Werkzeugen, damit diese leichter wiederverwendet werden können (cf. Burnard/Sperberg-McQueen 2002 und http://www.tei-c.org)

eines TEI-Dokuments des Schweizer Textkorpus mit Teilen des Header- und des Body-Teils mag den grundsätzlichen Aufbau des Formats verdeutlichen:

```
<?xml version="1.0" encoding="utf-8"?>
<TEI.2>
  <teiHeader>
  [...]
    <profileDesc>
      <textClass>
        <keywords>
          <term type="category">Belletristik</term>
          <term type="text_type">Prosaliteratur</term>
          [...]
        </keywords>
   [...]
   </teiHeader>
   <text>
     <body>
     [...]
     <l>Kommt VVFIN kommen 0</l>
     <l>,$,,0</l>
     <l>wir PPER wir 0</l>
     <l>fahren VVFIN fahren 0</l>
     <l>ein ART ein 0</l>
     <l>wenig PIS wenig 0</l>
     <l>hinaus PTKVZ hinaus 0</l>
     <l>. $. . 0</l>
     [...]
     </body>
   </text>
</TEI.2>
```

Im ersten Teil, dem Header, werden Metadaten in einer hierarchischen Struktur festgelegt. Im zweiten Teil, dem Body-Teil, folgt der eigentliche Text (z.B. <l>Kommt VVFIN kommen 0</l>) in der Form <l>Token Part-of-Speech-Tag Lemma Überschrift:ja/nein</l>[3].

Die Einhaltung dieses Formats ergibt sich bei einmal definierter Options-Datei gewissermassen automatisch, da sonst der Indexierprozess gar nicht ohne Fehler abschliessen würde.

Schwieriger, doch nicht weniger wichtig, ist die inhaltliche Vereinheitlichung der Formate. Wird ein Feld *Werkkategorie* definiert, so muss klar sein, welche Werte es annehmen kann; dies wiederum inhaltlich und formal einheitlich – formal darf das Feld nicht im einen Korpus den Wert 'Sachtext' annehmen, im anderen aber 'Sachtext**e**', inhaltlich muss 'Sachtext' in allen Korpora auf dieselbe Kategorie Texte angewandt werden.

3.2 Performance

Ein wichtiges Kriterium bei Webanwendungen allgemein ist die Performance bzw. sind die Antwortzeiten. Wer eine Webanwendung nutzt, wünscht sich ein flüssiges Navigationserlebnis und möglichst keine Wartezeiten.

Beim Korpus C4 wird naturgemäss am meisten Zeit für die Behandlung der Abfrage an die linguistische Suchmaschine verbraucht. DDC ist hier nach unseren bisherigen Erfahrungen sehr performant. Während bei informellen Tests in unserer Umgebung die DDC-Antwortzeit für unser lokal angebundenes Korpus für Einzelwortabfragen bei durchschnittlich 0.4 Sekun-

[3] Die Verwendung des Elements 'l' für ein Texttoken, die Anordnung 1 Token/Zeile sowie die Unterbringung der Zusatzinformationen zu den einzelnen Tokens tabulatorgetrennt im Inhaltsteil des l-Elements folgen nicht den TEI-Guidelines, sondern werden vom verwendeten Indexierer DDC so erwartet.

den lag, war dieser Wert bei einer verteilten Abfrage mit zusätzlich zwei entfernten Korpora bei immer noch guten 0.7 Sekunden.[4]

Die Effizienz in DDC rührt daher, dass einerseits die verschiedenen Korpora parallel abgefragt werden und dass andererseits das Zusammenführen der Treffer effizient gelöst ist. Gerade auf diese beiden Punkte müsste auch bei einer anderen Realisierung eines verteilten Korpusabfragesystems besonders geachtet werden. Denn wirklich durchsetzen wird sich nur ein System können, das von seiner Leistung her überzeugt.

3.3 Benutzerverwaltung

Mit der Benutzerverwaltung wird eher ein Randbereich eines verteilten Korpusabfragesystems angesprochen, der aber dann relevant wird, wenn wie beim Korpus C4 mehrere beteiligte Institutionen einen Hauptzugang zum verteilten Korpus anbieten wollen. Es stellt sich dann die Frage, ob die Benutzerdaten der registrierten Benutzerinnen und Benutzer geteilt werden, ob sie zentral gespeichert werden oder wie sonst mit ihnen verfahren wird.

Ganz ähnlich wie bei den Korpusdaten hat sich das Korpus C4 dafür entschieden, die Benutzerdaten getrennt zu halten, aber bei Bedarf verteilt abzufragen. Eine beim Korpus C4 in Basel registrierte Benutzerin z. B. kann das Korpus ganz normal vom Zugang in Basel aus benutzen. Sollte sie sich einmal beim entsprechenden Korpus-C4-Zugang in Wien anmelden, wäre sie dort im ersten Moment unbekannt, der Server würde aber gleich bei seinen Partner-Servern nachfragen, und da der Basler Server die Zugangsdaten als gültig rückmelden würde, könnte der Wiener Server die Nutzerin bei sich einloggen.

Dieses Vorgehen hat den Vorteil, dass, ausser punktuell beim Anmeldevorgang bei einem anderen Server, keine Nutzerdaten ausgetauscht werden. Ausserdem benötigen die Partnerserver kaum mehr Rechte für ein solch entferntes Einloggen als normale Korpusnutzerinnen und -nutzer. Das Vorgehen stellt damit einen einfachen und unkomplizierten Weg dar, Benutzerdaten gemeinsam zu verwalten, ohne dass die gesamten Datenbanken ausgetauscht werden müssen (wogegen wieder datenschutztechnische Gründe sprechen könnten).

4 Standardschnittstelle für Korpusabfragen

Zusammenschlüsse mehrerer Korpora bzw. die Nutzung mehrerer Korpora über verteilte Abfragen sollte nicht auf das Korpus C4 beschränkt bleiben, sondern wäre, wie in Abschnitt 2 angesprochen, in vielen Situationen wünschenswert. Um verschiedene Korpora verteilt abfragbar zu machen, ist eine gemeinsame Schnittstelle zwingend notwendig. Damit möglichst viele Korpusprojekte möglichst einfach an verteilten Korpusabfragesystemen teilnehmen können, sollten für Schnittstellen für Korpusabfragen Standards erarbeitet werden.

Das Korpus C4 selber ist bei seiner Realisierung eines verteilten Textkorpus stark nach Bottom-up-Ansätzen verfahren, indem vorerst nur die verteilte Abfrage in einer einfachen Version nur unter DDC realisiert wurde. Allgemeinere Standardschnittstellen sind dabei nicht definiert worden. Im Folgenden sollen die Eigenschaften einer Standardschnittstelle für Korpusabfragen nur skizziert werden, jedoch mit besonderer Berücksichtigung der Erfahrungen und Probleme aus dem Projekt Korpus C4.

[4] Abfragen mit jeweils 100 zufällig ausgewählten Wörtern unterschiedlicher Häufigkeitsklassen; maximal 50 zurückgegebene Treffer.

4.1 Voraussetzungen

4.1.1 Offene Standards und offene Software

Eine wichtige Voraussetzung für eine Standardschnittstelle ist ihre Zugänglichkeit. Um sie so zugänglich wie möglich zu gestalten, sollte wo immer möglich auf offene Standards und offene Software gesetzt werden. Die Verwendung von Standards ist in der Korpuslinguistik bereits gut verankert, man denke zum Beispiel an TEI (cf. Burnard/Sperberg-McQueen 2002) mit seiner weiten Verbreitung oder De-Facto-Standards wie das STTS-Tagset beim Part-of-Speech-Tagging für das Deutsche[5]. Die grosse Akzeptanz von Standards ist zugleich ein Argument dafür, dass auch ein Schnittstellenstandard gut aufgenommen werden könnte.

Gut zugängliche Werkzeuge, d. h. in vielen Fällen quelloffene Software, sind ein weiterer wichtiger Punkt. Zusammen mit Standards bilden sie die Basis für Kooperationen zwischen Korpusprojekten, wie sie zum Aufbau verteilter Korpora nötig sind. Zudem verlangen die Werkzeuge selber (Part-of-Speech-Tagger, Indexierer etc.) oft die Einhaltung gewisser Standards und Formate. Die gute Verfügbarkeit eines bestimmten Werkzeugs und ein leichter Zugang dazu kann so ein starkes Argument für die Verwendung eines zugehörigen Standards sein.

4.1.2 Systemvielfalt

Die meisten Korpusprojekte sind nicht für die verteilte Nutzung konzipiert worden, und dies wird wohl auch in Zukunft so bleiben. Je kleiner der Aufwand ist, ein Korpus in einen Verbund einzupassen, desto grösser ist die Chance, dass es tatsächlich geschieht. Der Idealfall wäre, das Korpusprojekte dasselbe Korpus unverändert für ihre eigenen Zwecke wie auch für verteilte Abfragen nutzen könnten.

Das bedingt jedoch, dass eine Standardschnittstelle mit einer möglichst grossen Systemvielfalt umgehen kann. Beim Korpus C4 war diese Systemvielfalt ausgeschaltet worden (cf. 3.1). Es zeigte sich aber, dass die Abstimmung der Systeme und ihre Vereinheitlichung einen beträchtlichen – vor allem auch koordinatorischen – Aufwand nach sich zog.

4.2 Aufgabenverteilung

Abbildung 2 illustriert schematisch die Aufgabenverteilung der verschiedenen Mitspieler.

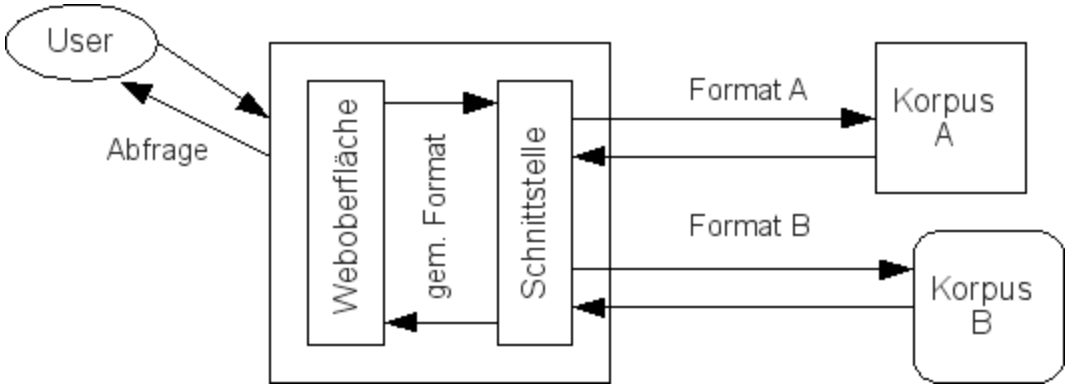

Abb. 2: Aufgabenverteilung Schnittstelle – Teilkorpora

Aufgabe der Standardschnittstelle ist es, zwischen einer Abfrageoberfläche, mit der die Nutzerinnen und Nutzer interagieren, und den einzelnen verteilt abgefragten Korpora zu

[5] STTS steht für 'Stuttgart-Tübingen Tagset', ein Part-of-Speech-Tagset, entwickelt von den Universitäten Stuttgart und Tübingen (cf. Schiller et al. 1995). Das Tagset ist wohl das meistverwendete für das Deutsche und wird auch beim Schweizer Textkorpus und dem Korpus C4 eingesetzt.

vermitteln. Die Abfrageoberfläche nimmt Abfragen in ihrem Format entgegen und erwartet Resultate in ihrem Format zurück (idealerweise gleich das Format der Schnittstelle).

Die Korpora auf der anderen Seite bieten ihre Dienste ebenfalls in ihrem eigenen Format an und antworten auf Abfragen nur in ihrem eigenen Format.

Die Schnittstelle nun übersetzt Abfragen der Abfrageoberfläche ins Format bzw. in die Formate der einzelnen Korpora, gibt die übersetzten Abfragen weiter, nimmt die Resultate in den jeweiligen Formaten der Korpora entgegen, übersetzt sie zurück ins Format der Abfrageoberfläche und reicht sie an diese weiter.

Um den Aufwand auf Seiten der Teilkorpora zu minimieren, sollte die Schnittstelle möglichst alle Übersetzungsleistungen übernehmen, so dass ein abgefragtes Teilkorpus keine zusätzlichen Dienste anbieten muss als jene, die es bereits von Haus aus anbietet.

4.3 Schnittstellendefinition

Da sich analoge Problemstellungen auch in anderen Gebieten zeigen, kann für die Definition der Schnittstelle auf dortige Vorarbeiten zurückgegriffen werden. Namentlich im Bereich der Sensornetzwerke und verteilten Systeme existiert mit den *Global Sensor Networks (GSN)* (cf. Aberer/Hauswirth/Salehi 2006) ein Ansatz, der als Modell für die hier propagierte Standardschnittstelle für Korpusabfragen dienen könnte.

GSN definiert ebenfalls eine Schnittstelle für die gemeinsame Abfrage verteilter Datenquellen unterschiedlichen Zuschnitts. Mittels GSN werden Sensoren und Sensornetzwerke (also z.B. Bewegungsmelder, Lichtsensoren, Webcams mit Netzanbindung) miteinander verbunden, gemeinsam abgefragt und die Resultate konsolidiert dargestellt. Die Schnittstelle muss dabei fähig sein, gleichzeitig mit ganz unterschiedlichen Systemen zu kommunizieren.

In einer XML-Konfigurationsdatei werden bei GSN die beteiligten Sensoren als virtuelle Sensoren modelliert. Das Frontend benutzt dabei diese virtuellen, vereinheitlichten Sensoren. In der Beschreibung des virtuellen Sensors in der Konfigurationsdatei wird eine Abstraktionsstufe tiefer angegeben, auf welche Weise auf welche Daten des Sensors zugegriffen werden kann. Das gekürzte Beispiel der Definition eines einfachen Temperatursensors aus Aberer/Hauswirth/Salehi (2006: 7) mag dies illustrieren.

```
<virtual-sensor name="Light-sensor1" priority="11">
  [...]
  <description>A TinyOS temperature vsensor</description>
  [...]
  <output-structure>
    <field name="temperature" type="int" />
  </output-structure>
  <storage history-size="10s" permanent-storage="true" />
  <input-streams>
    <input-stream name="temperature" >
       <stream-source alias="tsensor" storage-size="1">
         <address wrapper="tinyos">
           <predicate key="host">lsirpc24.epfl.ch</predicate>
           <predicate key="port">9001</predicate>
         </address>
         <query>
           select WRAPPER.TEMPERATURE as temperature,
                  WRAPPER.TIMED as timestamp from WRAPPER
         </query>
       </stream-source>
       <query>
          select temperature from tsensor
       </query>
     </input-stream>
   </input-streams>
</virtual-sensor>
```

Es wird definiert, welchen Output von diesem Sensor zu erwarten ist (*output-structure*), nämlich ein Feld *temperature* als Ganzzahl. Woher und in welcher Form der Datenstrom kommt, wird unter *stream-source* festgelegt: Die Adresse der Datenquelle, über welchen Wrapper (*tinyos*) sie anzusprechen ist und welche Datenbankabfrage (1. *query*) zum gewünschten Wert führt.

Ganz ähnlich könnte man sich die Definition einer Standardschnittstelle für Korpusabfragen vorstellen. Eine Abstraktionsschicht auf der Schnittstelle würde virtuelle Korpora definieren. Darin würde beschrieben, welche Daten das Korpus in welcher Form anbiete, über welche Adresse, mit welcher Methode und welchen konkreten Abfragen darauf zugegriffen werden kann.

Eine Abstraktionsebene höher kann dann einheitlich auf diese virtuellen Korpora zugegriffen werden. Ein verteiltes Korpus schliesslich würde selber definiert als ein komplexes virtuelles Korpus, das mehrere virtuelle Korpora vereint. Eine solche Konfiguration ähnelt stark dem Aufbau des Korpus C4: Ein Verbund gleichartiger Korpora als ein Korpus – lediglich eine Abstraktionsebene höher als beim jetzigen Korpus C4.

4.3.1 Wrapper

Bestandteil der Schnittstelle müssten Wrapper sein, welche die Aufgabe übernehmen, die Abfragen, wie sie von der Abfrageoberfläche her kommen, zu übersetzen, so dass sie für das angesprochene Korpus verständlich werden. Mit den zurückerhaltenen Resultaten müsste die Übersetzung in umgekehrter Richtung ebenfalls durch diesen Wrapper erfolgen.

Man hätte dann z. B. einen Wrapper für DDC, der eine Abfrage vom gemeinsamen Format ins DDC-Format übertragen, diese Query anschliessend für DDC verständlich (als Socket-Message) absetzen und das Resultat entgegennehmen und ins gemeinsame Format zurückübersetzen würde.

4.3.2 Abfrageoberfläche

Abfrageoberflächen für diese Standardschnittstelle, welche die Abfragen von Benutzerinnen und Benutzern annehmen und Resultate am Bildschirm präsentieren, würden sich kaum von den heutigen unterscheiden. Der Hauptunterschied wäre, dass sie gegen die Standardschnittstelle programmiert wären und nicht gegen einen der spezifischen Korpusserver.

4.4 Gemeinsame Daten

Im Korpus C4 war eines der grösseren Hindernisse das Finden der gemeinsamen Daten und Metadaten. Eine Möglichkeit, dieses Problem etwas abzumildern und die Charakteristika der Teilkorpora in der gemeinsamen Abfrage nicht ganz zu verlieren, könnte sein, eine gewisse Asymmetrie zwischen Suche und Darstellung einzuführen (es kann nur nach Gemeinsamem gesucht werden, dargestellt wird aber "alles").

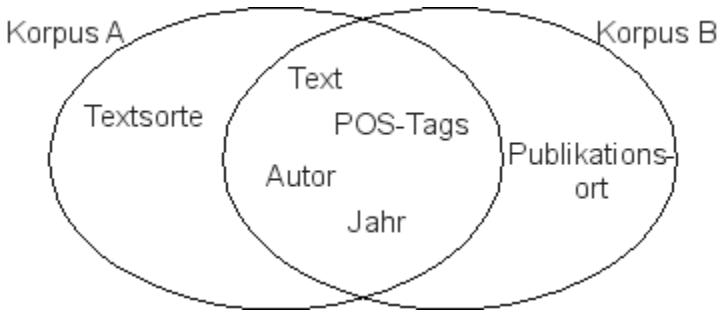

Abb. 3: Zwei Korpora mit gemeinsamen und unterschiedlichen Daten

Abbildung 3 zeigt vereinfacht zwei Korpora mit teils gemeinsamen, teils verschiedenen Daten und Metadaten. Von der Abfrageoberfläche her sinnvoll abgefragt werden können nur Daten, die beiden beteiligten Korpora gemein sind. Wenn wir also beide Korpora gemeinsam abfragen wollen, können wir nur die Schnittmenge der definierten Daten nutzen.

Etwas anders liegt der Fall bei der Darstellung der gefundenen Treffer. Hier kann es je nachdem sinnvoll sein, mehr als nur die gemeinsamen Daten anzuzeigen. Im Beispiel aus Abbildung 3: Obwohl wir nicht nach der Textsorte suchen können, möchten wir bei den Treffern aus Korpus A die Textsorte vielleicht doch angezeigt bekommen. Prinzipiell können für die Anzeige alle verfügbaren Felder aller Korpora genutzt werden. Bei grossen Abweichungen kann dies jedoch schnell zu Felderwildwuchs führen und schwer übersichtlich darstellbar werden.

4.5 Metadatendefinitionen und Zuordnungstabellen

Es reicht nicht aus, gemeinsame Datenfelder zu definieren, es muss auch sicher gestellt sein, dass sie einander über die Korpusgrenzen hinweg inhaltlich und formal entsprechen. Beim Korpus C4 wurde dies so gelöst, dass alle Teilprojekte ihre gemeinsam angebotenen Daten vereinheitlicht haben (cf. Abschnitt 3.1.2). Für den allgemeinen Fall, wo alle Teilkorpora ihre Daten möglichst unverändert anbieten können sollen, ist dies keine Option.

Das Problem ist, ausser es handle sich um völlig inkombatible Kategorien, wieder ein Formatkonvertierungsproblem, ähnlich demjenigen der unvereinbaren Formate der verschiedenen Korpusserver. Die Lösung könnte denn auch ganz ähnlich aussehen wie die dafür propagierten Wrapper (cf. Abschnitt 4.3.1). In diesem Stadium wären die Daten bereits im eigenen Format der Standardschnittstelle, so dass es nur noch um die Umsetzung von Zuordnungen ginge. Die Vereinheitlichung von Kommunikationswegen und Low-Level-Formaten fiele hier weg, da schon durchgeführt. Es müsste lediglich festgelegt werden,

welche Werte der Felder eines Teilkorpus welchen Werten im Gesamtkorpus entsprechen. Eine Zuordnungstabelle, die diese Beziehungen definiert, könnte etwa folgendermassen aussehen:

```
[...]
<mappings corpus="Korpus A">
  <mapping field="Werkkategorie">
    <relation>
      <value>Sachtext</value>
      <mapped_to>Sachtexte</value>
    </relation>
    <relation>
      <value>Wissenschaftlicher Text</value>
      <mapped_to>Sachtexte</value>
    </relation>
    [...]
  </mapping>
  [...]
</mappings>
```

Das Gesamtkorpus definiert in diesem Beispiel ein Feld *Werkkategorie*, das den Wert 'Sachtexte' annehmen kann. Korpus A hingegen kennt unter *Werkkategorie* die Werte 'Sachtext' und 'Wissenschaftlicher Text', die inhaltlich dem Wert 'Sachtexte' im Gesamtkorpus entsprechen.

Diese Zuordnungstabellen wären in den meisten Fällen projektspezifisch. Lediglich für oft gebrauchte und standardisierte Metadatenzuordnungen, wie etwa der Vermittlung zwischen zwei unterschiedlichen Part-of-Speech-Tagsets, könnte man sich eine breitere Wiederverwendung vorstellen.

Der Gebrauch von Standards bei der Kodierung digitaler Korpustexte, gemeint ist vor allem TEI, kann dabei eine grosse Hilfe sein: Die Bandbreite möglicher Felder wird dabei eingeengt, so dass das Vorhandensein kompatibler Felder wahrscheinlicher wird. Ausserdem sieht TEI (mindestens implizit via DTD, Schema etc.) die Dokumentation der verwendeten Felder vor – eine Grundvoraussetzung zur Herstellung von Zuordnungen zwischen Feldern.

4.6 Umsetzung

Konkrete Pläne für eine Realisierung der hier skizzierten Ideen zu einer Standardschnittstelle für Korpusabfragen existieren bislang nicht. Aus unserer Erfahrung heraus ist es aber sicher empfehlenswert, dabei mit einem bestehenden, funktionierenden System zu beginnen, also z. B. mit dem Korpus C4, und schrittweise darauf aufzubauen. Die hier beschriebene Standardschnittstelle ist auch genügend modular aufgebaut, so dass problemlos einzelne Teile (z. B. Metadatenmapping, Wrappen verschiedener Korpusserver etc.) relativ unabhängig von anderen erstellt werden können. Ungereimtheiten könnten dabei gleich mit Praxisbezug ausgeräumt werden, und noch nicht vollständig Ausdefiniertes müsste dabei natürlich exakt spezifiziert werden.

5 Schluss

Mit dem Korpus C4 steht jetzt ein verteilt abfragbares Korpus für das Deutsche zur Verfügung. Es ist zwar noch weit davon entfernt, eine Standardschnittstelle für Korpusabfragen zu definieren. Als voll funktionsfähiges verteiltes Korpus kann es aber den Nutzen dieser Art Korpus aufzeigen und zur Nachahmung animieren. Es kann zudem als Ausgangspunkt und Referenz dienen für die Realisierung einer Standardschnittstelle für Korpusabfragen wie sie in Abschnitt 4 propagiert wird. Die Erfahrungen aus dem Korpus C4 können bei der genaueren Spezifikation helfen, die für die hier erst grob skizzierten

Eigenschaften der Standardschnittstelle sicher noch nötig würde. Ausserdem kann bei einer Realisierung immer wieder mit dem Korpus C4 als theoretisch einfacherer Variante eines verteilten Korpus direkt verglichen oder gar direkt auf ihm aufgebaut werden. Es ist auf jeden Fall zu hoffen, dass sich die verteilte Nutzung linguistischer Korpora in Zukunft noch weiter ausbreiten wird.

Literatur

Aberer, Karl/Hauswirth, Manfred/Salehi, Ali (2006): *Global Sensor Networks*. Lausanne: Ecole Polytechnique Fédérale de Lausanne (EPFL), Tech. Rep. LSIR-REPORT-2006-001.

Bickel, Hans et al. (2009): "Schweizer Text Korpus". Erscheint in *Linguistik online*.

Burnard, Lou/Sperberg-McQueen, Christopher Michael (eds.) (2002): *Guidelines for Electronic Text Encoding and Interchange*. Oxford: Humanities Computing Unit, University of Oxford.

Cunningham, Hamish et al. (2002): "GATE: A framework and graphical development environment for robust NLP tools and applications". In: Association for Computational Linguistics (ed.): *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania.

MacWhinney, Brian (1991): *The Childes Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates.

Neuroth, Heike/Kerzel, Martina/Gentzsch Wolfgang (eds.) (2007): *Die D-Grid Initiative*. Göttingen: Universitätsverlag Göttingen.

Quasthoff, U./Richter, M./Biemann, C. (2006): "Corpus Portal for Search in Monolingual Corpora". In: *Proceedings of the LREC 2006*. Genoa, Italy.

Schiller, Anne/Teufel, Simone/Thielen, Christine (1995): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Stuttgart, Tübingen: Institut für maschinelle Sprachverarbeitung, Universität Stuttgart; Seminar für Sprachwissenschaft, Universität Tübingen.

Sokirko, Alexey (2005): *A Technical Overview of DWDS/Dialing Concordance*. http://www.aot.ru/docs/OverviewOfConcordance.htm, Stand Juni 2009.