SPre fix

Anleitung zum Arbeiten mit SPre unter GATE

Jürgen Hermes 31.12.2004

SPre ist ein Programm, mit denen Texte beliebigen Formates segmentiert und annotiert werden können. Die Algorithmen zur Segmentierung sind mittels einer XML-Datei relativ frei konfigurierbar. Ebenso können eigene Annotatoren in das Projekt integriert werden. Das SPre-Projekt ist demnach so konstruiert, dass es eine grundlegende Architektur bereitstellt, in die Komponenten als Plugins eingestöpselt werden können. Diese Anleitung wurde indess für Nutzer geschrieben, die ohne lange Einarbeitungszeit einen Präprozessor für die Auszeichnung von Sprachdaten verwenden wollen. Deshalb findet sich hier keine Anleitung zum Programmieren eigener Komponenten, sondern lediglich Beschreibungen zur Verwendung der vorhandenen. Alles weitere findet sich im ausführlichen Manual (Erscheint voraussichtlich im Februar 2005). Wir haben uns entschieden, SPre zunächst als Plugin für das Projekt GATE zu veröffentlichen, da die graphische Oberfläche ausgereift ist und eine Fülle von Möglichkeiten zur Visualisierung bietet die eventuell schon dem ein oder anderen Nutzer bekannt sein dürfte.

Zur einfachstmöglichen Nutzung von SPre unter GATE müssen Sie die folgenden Schritte unternehmen:

- Laden Sie sich das GATE-System vom Server der Universität Sheffield herunter (http://gate.ac.uk). Eine gut lesbare und zuweilen spaßige Einführung in GATE bietet das sich ebenfalls auf dieser Seite befindliche Tutorial (http://gate.ac.uk/sale/tao/index.html). Folgen Sie den darin enthaltenen Anweisungen zu Download und Installation. Alles weitere, was das Arbeiten mit SPre unter GATE betrifft, finden Sie in dieser Anleitung.

- Laden Sie sich das SPre-Plugin vom Server der Sprachlichen Informationsverarbeitung herunter, entweder als komplettes Projekt (http://spinfo.uni-koeln.de/forschung/spre/spre.zip) oder als binäres *jar*-Archiv, das mit den benötigten Konfigurationsdateien und Ressourcen in eine Datei gepackt wurde (http://spinfo.uni-koeln.de/forschung/spre/spre_small.zip).

- Haben Sie sich für die binäre Version entschieden, so entpacken Sie die Datei *spre_small.zip* in ein Verzeichnis *SPre*, so erhalten Sie den Ordner *misc*, der die Steuerdatei für den Präprozessor (*SPreConfig.xml*), einige eventuell benötigte Ressorcen (Abkürzungsdatei etc.) sowie einen Unterordner *plugouts* enthält. In diesem findet sich der Ordner *gate*, der wiederum das Projekt als Archiv (*spre.jar*), sowie Informationen über SPre für Gate (*creole.xml*) enthält. Wenn Sie die Vollversion heruntergeladen haben, entpacken Sie die Datei *spre.zip* in ein Verzeichnis, so erhalten Sie neben dem soeben beschriebenen *misc*-Verzeichnis noch die Verzeichnisse *src*, *bin*, *doc* und *lib*. *src* enthält den Quelltext des Projektes SPre, *bin* die Binärdateien, *doc* die javadoc-Dokumentation und *lib* die benötigten externen Archive.

- Starten Sie GATE, so erscheint eine graphische Oberfläche, die im Groben aus zwei Teilen besteht: Auf der linken Seite befindet sich ein schmalerer Frame, in dem sich vier Symbole für unterschiedliche verwendbare Komonenten befinden (*Applications*, *Language Ressources*, *Processing Ressources*, *Data Stores*). Rechts befindet sich ein Frame, auf dem bisher erst ein einziges Registerblatt (*Messages*) liegt. Über beiden Frames finden sich Menuepunkte.

- Zunächst müssen Sie Gate nun mitteilen, wo sich die Komponente befindet, die Sie nutzen wollen.

Klicken Sie dafür auf den Menüpunkt *File* → *Manage Creole Plugins*. Es öffnet sich ein neues Fenster, in dem die bisher dem GATE-System bekannten Plugins aufgeführt sind. Sie können das SPre-Plugin hinzufügen, indem Sie über die Schaltfläche *Add a new CREOLE repository* einen Dialog aufrufen, in dem Sie den Pfad zu dem neuen Plugin angeben. Sie müssen in unserem Fall hier den Ordner *SPre/misc/plugouts/gate* angeben, mit *OK* bestätigen. Alsdann müssen Sie bestätigen, daß Sie die in diesem Ordner spezifizierten Komponenten (SPre) laden wollen (entweder nur für diese Session oder bei jedem Start von GATE).

✗ Nun können Sie eine Instanz von SPre als neue Processing Resource (PR) anlegen. Dafür klicken Sie mit der rechten Maustaste auf das Symbol *Processing Resources* im linken Frame. Gehen Sie auf das erscheinende *New* und wählen Sie aus den erscheinenden Komponenten *SPre* aus. Es erscheint ein Dialog, in dem Sie der neuen Komponente einen Namen geben können und den Pfad zur SPre-Konfigurationsdatei angeben müssen (weitere Pfade brauchen Sie vorerst nicht zu setzen, s.u.). Diese findet sich – wie oben bereits erwähnt – im entpackten Verzeichnis *SPre.misc* (*SPreConfig.xml*).

✗ Jetzt brauchen Sie noch einen Text, den Sie prozessieren wollen. Ein solcher Text wird in GATE als Language Resouce (LR) eingebettet. Sie können entweder einen eigenen oder auf den sich ebenfalls in *SPre.misc* befindlichen Beispieltext (*faz1.xml*) angeben. Dafür machen Sie einen Rechtsklick auf das Symbol *Language Resources* und wählen unter *New* den Punkt *GATE Document* aus. Im erscheinenden Dialog können Sie fakultativ Namen und weitere Parameter für diese LR angeben, obligatorisch ist die Angabe des Pfades zur zu prozessierenden Datei.

✗ Nun müssen Sie eine Applikationsinstanz schaffen, in welcher ausgewählte Komponenten verknüpft und ausgeführt werden. Dazu machen Sie wiederum einen Rechtsklick, diesmal auf *Applications* und wählen unter *New* den Punkt *Pipeline* aus. Geben Sie im folgenden Dialog der neuen Pipeline einen Namen und bestätigen Sie mit *OK*.

✗ Ein Doppelklick auf die neu erstellte Applikation öffnet eine neue Registerkarte im Frame rechts. In dieser markieren Sie ihre vorhin erstellte PR und betägigen den Pfeil nach rechts. Die PR ist nun ausgewählt. Klicken Sie auf die ausgewählte Komponente, so können Sie eine ihrer instanziierten LR im unteren Feld (*Parameters for the [zugewiesener Name] SPre*) zur Prozessierung auswählen. Ein Klick auf den Button *Run* rechts unten startet die Prozessierung.

✗ Erfolgte die Prozessierung mit den Voreinstellungen in der mitgelieferten Konfigurationsdatei. Wenn die dort codierten Pfade zu Ressourcen (hier: Datei mit Abkürzungen) nicht mit denen auf ihrem Rechner übereinstimmen und Sie auch keine weiteren Pfade in der Instanz der PR eingegeben haben, so erfolgte die Prozessierung ohne Rückgriff auf Ressourcen. Dies wird Ihnen auf der *Messages*-Registerkarte im rechten Frame angegeben. Wenn Sie dies ändern wollen, so müssen Sie entweder in der SPre-Konfigurationsdatei oder in der PR-Instanz (letztere überschreibt eventuell vorgenommene Einträge in der Konfigurationsdatei) die Pfade zu den betreffenden Ressourcen setzen.

✗ Sie haben SPre nun mit der von uns voreingestellten Konfiguration auf einen von Ihnen ausgewählten Text angewendet. Das Ergebnis können Sie sich auf zweierlei Arten anschauen (es existieren genaugenommen noch mehr Möglichkeiten, die hier nicht aufgeführt sind):

- indem Sie ihren Text doppelklicken. Er erscheint nun auch als Registerkarte im rechten Frame. Wählen Sie das Schaltfeld *Annotation Sets* aus und klicken sie auf den unbenannten Pfeil. Die Auszeichnungen, welche von SPre im Text vorgenommen wurden, erscheinen in unterschiedlichen Farben. Sie können nun einzelne Elemente auswählen. Diese werden im Text markiert.

- indem Sie in Ihr Arbeitsverzeichnis gehen und dort die Datei *output.xml* aufrufen, die vom

voreingestellten Persistor erzeugt wurde. Auch der Pfad zu dieser Ausgabedatei kann sowohl in der Konfigurationsdatei als auch in der PR-Instanz verändert werden.

Sie können nun beliebig viele weitere Texte prozessieren und eventuell auch in der Konfigurationsdatei *SPre/misc/SPreConfig.xml* Änderungen vornehmen, um zu überprüfen, welche Auswirkungen das auf die Prozessierung hat. Die Konfigurationsdatei ist intern relativ ausführlich kommentiert, so dass hier auf die einzelnen Komponenten nicht näher eingegangen werden muss. Mit den Default-Einstellungen wird der Input lediglich segmentiert. Um eine Beispiel-Annotationskomponente einzubinden, müssen Sie den auskommentierten mitgelieferten Annotator *DistributionalMorphemizer* in der Konfigurationsdatei aktivieren. Dann benötigen Sie allerdings noch die Ressource *morphemes.zip*, die Sie eigens herunterladen müssen (siehe http://spinfo.uni-koeln.de/forschung/spre/resources/). Falls Sie eigene Komponenten für SPre schreiben und einbinden wollen, so sind Sie entweder auf das Studium der bisher leidlich kommentierten Quelltexte angewiesen oder müssen sich noch etwas Geduld heben, bis das vollständige Manual erscheint (wir gehen von Mitte Februar 2005 aus).