TextPrep

ein ToolSet zur

Erzeugung von Wortlisten
und zur
Fehlerbereinigung von Quelltexten
fürdas
***Dragon* NaturalVocTool**

zur Erstellung von Fachvokabularen
für
***Dragon* NaturallySpeaking**
von
Nuance

>>> Kurzbeschreibung <<<

Dr. Tilmann Bäßler

März 2007

Ein Produkt der Bein-EDV

Im Benutzerhandbuch des Dragon NaturalVocTools wird eindringlich darauf hingewiesen, dass die Quelltexte zur Erstellung eines Fachvokabulars aufbereitet sein müssen, bevor sie als Input für das NaturalVocTool verwendet werden können.

Neben der Entfernung von Bildern, Tabellen und allen Formatierungsinformationen ist es vor allem wichtig,

- dass der Text fehlerfrei ist,
- dass man Wortlisten erstellt hat, z.B.:
- Abkürzungen mit ihrergesprochenen Form,
- Fachausdrücke, die aus mehreren Worten bestehen,
- mit Bindestrich verketteten Begriffe, bei denen der Bindestrich nicht gesprochen werden muss (z. B. ,,Mutter-Kind-Beziehung“),
- Ausdrücke mit einem sog. Ergänzungsstrich (z.B. ,,Hals- und
- Beinbruch“, oder ,,Ankunftszeit und -ort“)
- Maßeinheiten und sonstige Sonderworte mit ihrer gesprochenen Form.

Was leistet nun das Toolset TextPrep?

A) Quelltextbearbeitung

- Zeigt alle Wörter, Abkürzungen, usw. in ihrer Textumgebung, um ihre Bedeutung besser verstehen zu können.
- Zeigt den Quelltext in seiner gedruckten Form und parallel dazu die darin vorkommenden Sonderzeichen und nicht druckbaren Zeichen in ihrer Hexadezimalform. Dies erleichtert das Auffinden besonderer Zeichen bzw. Wortgebilde.
- Erlaubt, alle Dokumente mit Sonderzeichen und nicht druckbaren Zeichen von anderen Quelltexten zu trennen, um sie in einem nachfolgenden Schritt einheitlich bearbeiten zu können. So können z.B. alle Texte mit DOS ASCII-Zeichen bearbeitet werden, damit sie unter Windows ANSI die Umlaute korrekt darstellen.
- Es ist wichtig, Quelltexte in Gruppen von gleichem Textaufbau zu unterteilen, da dadurch störende Adress- und Namenselemente leichter entfernt werden können. TextPrep trennt auf einfache Weise die Quelldokumente nach ihrer Struktur anhand des Dokumentanfangs oder anhand von Ausdrücken innerhalb des Dokuments.
- Manche Textkorrekturen müssen beim Erstellen von Fachvokabularen immer wieder durchgeführt werden. TextPrep merkt sich diese Korrekturen und führt sie ,,auf Knopfdruck“ bei jedem neuen Quelltext erneut aus. So gelingt es z.B., einheitliche Abkürzungen anzubieten, oder Quelltexte aus der DOS-Zeit in Windows ANSI Texte umzusetzen.
- Durch Worttrennungen am Zeilenende entstehen im VocTool viele unerwünschte Wortfragmente. TextPrep fügt diese Wortteile wieder zusammen und ,,rettet“ diese und ihre Wortumgebung für das Vokabular.
- Benennt alle Quelldateien systematisch um, um evtl. Hinweise auf Patienten zu eliminieren.
- Fasst auf Wunsch alle Quelltextdateien zu 1 oder mehreren großen Dateien zusammen.

- Ist dann schließlich der Quelltext mit TextPrep bearbeitet, ruft man das Nuance VocTool auf, um eine erste Liste aller unbekannten Worte zu erzeugen. Man stellt in aller Regel fest, dass doch noch sehr viele Wörter mit Schreibfehlern vorhanden sind, die dann mühsam Wort für Wort mit Hilfe von z.B. MS Word und der Funktion ,,Ersetzen“ korrigiert werden müssen, um den Kontext für die Bi- und Trigrammstatistiken zu erhalten und ein fehlerfreies Vokabular nach einem zweiten Lauf von VocTool anbieten zu können. Auch hier erleichtert TextPrep das Umsetzen der Korrekturen: Mittels TextPrep bearbeitet man die Wortliste von VocTool mit einer Reihe von Hilfsmitteln, bis die Schreibweise der einzelnen Begriffe korrekt ist. Und damit ist die Arbeit auch schon getan: TextPrep überträgt diese Änderungen selbsttätig auf den Quelltext!
- Weitgehendes Umsetzen von Quelltexten alter deutsche Rechtschreibung in neue deutsche Rechtschreibung!
- Löscht unerwünschte Textabschnitte (Adressteile, Personennamen, Literaturstellen, usw.) am Anfang, am Ende, und beliebig auch mitten in den Quelltextdokumenten.

B) Erstellen von Wortlisten, z.T. mit der gesprochenen Form

- Erstellt eine Liste allerAbkürzungen. In einem zweiten Schritt kann aus früheren Abkürzungslisten die gesprochene Form übertragen werden.
- Erstellt eine Liste aller Sonderbegriffe mit mehr als 1 Großbuchstaben (z.B. "ADAC", "MHz") einschließlich der gesprochenen Form.
- Erstellt eine Liste aller Begriffe, die keinen Vokal enthalten. Dies sind sehr häufig Maßeinheiten (z.B. mg/ml, µg, °C) einschließlich der gesprochenen Form.
- Erstellt eine Liste aller Mehrwortbegriffe (z. B. ,,Mutter-Kind-Beziehung“) einschließlich der gesprochenen Form (d.h. ohne Bindestriche), damit die Bindestriche nicht mitdiktiert werden müssen. Neben dem Bindestrich lassen sich auch andere Trennzeichen wie z.B. ,,@“ auswählen, um so eine Liste aller Internetadressen zu erzeugen.
- Erstellt eine Liste aller Begriffe, die einen Ergänzungsstrich beinhalten z.B. ,,Hals- und Beinbruch“ oder ,,Knochenfrakturen und –behandlungen“ einschließlich der gesprochenen Form (d.h. ohne Bindestriche).
- Erstellt eine Liste von aufeinanderfolgenden Begriffen, die groß geschrieben sind. Auf diese Weise werden z.B. Namen wie "Deutsches Rotes Kreuz" oder "Verband Deutscher Elektriker" gefunden.
- Adjektive oder Verben, die als Substantive gebraucht werden, werden von Dragon NaturallySpeaking oft nicht als solche erkannt und daher irrtümlich klein geschrieben. Beispiele: "beim Spielen", "etwas Schönes", "nach Abdecken und Abwaschen". TextPrep erzeugt eine Wortliste mit diesen Begriffen, um die korrekte Großschreibung sicherzustellen.
- TextPrep vervollständigt Fachausdrücke, deren 1. Wort in einer Liste angegeben ist.

  Beispiel: Es sind vorgegeben: "Arcus", "Canalis", "Corpus".
  Diese ergänzt TextPrep ganz oder teilweise zu
  "Arcus plantaris pro fundus", "Canalis semicularis", "Corpus adiposum pararenale", sofern diese im Fließtext vorhanden sind. Die so ergänzten Ausdrücke können nun dem Vokabular als eigenständige Ausdrücke hinzugefügt werden.

Kurzbeschreibung der einzelnen Tools

CONTEXT

Alle Dateien im Quellverzeichnis werden nach einer Zeichenkette durchsucht. Die zu suchende Zeichenkette wird erfragt. Pro Datei wird wahlweise die erste Fundstelle mit Kontext angezeigt (pro Datei 1 Zeile) oder es werden alle Fundstellen angezeigt. Dieses Tool ist hilfreich, um Abkürzungen und Sonderbegriffe in ihrer Bedeutung im Kontext zu verstehen bzw. bei der Bearbeitung von Wordlisten automatisch zu sehen, in welchem Kontext der markierte Ausdruck vorkommt.

SHOWHEX

Einzelne Texte im Quellverzeichnis können betrachtet werden. Dabei werden alle Sonderzeichen hexadezimal dargestellt.

Mit diesem Tool erkennt man, ob z.B. die Umlaute nicht im erforderlichen ANSI TXTFormat vorliegen, sondern im DOS ASCII-Format.

WRONGHEX

Dieses Tool arbeitet eng mit SHOWHEX zusammen. Alle in SHOWHEX gefundenen Dokumente, die ,,falsche“ Hex-Zeichen enthalten, können mit WRONGHEX extrahiert werden, damit in einem nachfolgenden Schritt (mit dem Tool REPLACE) diese Zeichen in die ,,richtigen“ ANSI-Zeichen umgesetzt werden können.

TXTSTART

Ein wichtiges Element der Dokumentaufbereitung ist die Anonymisierung der Quelltexte: Das zu erstellende Fachvokabular sollte weder Personennamen noch sonstige Adresselemente enthalten. Viele Quelltexte sind in ihrer Struktur einheitlich aufgebaut. Sie beginnen z.B. häufig mit dem Briefkopf.

Das Tool TXTSTART erlaubt, die Quelltexte nach ihrer Struktur zu trennen. Dateien mit einheitlicher Struktur können dann in einem nachfolgenden Schritt (mit dem Tool DELLINES) so bearbeitet werden, dass die Namens- und Adresselemente entfernt werden können.

REPLACE

Auf alle Dateien im Quellverzeichnis werden Ersetzungsvorschriften (d.h. Textkorrekturen) angewendet, die in der zugehörigen Tabelle festgelegt wurden. Alle Dateien des Quellverzeichnisses werden nach dem Ersetzungsvorgang im Zielverzeichnis gespeichert.

Dieses Tool kann sehr vielfältig eingesetzt werden. Es gibt gleich bleibende Ersetzungen, die beim Erstellen neuer Fachvokabulare jedes Mal vorgenommen werden müssen. Da bei TextPrep diese Ersetzungsanweisungen gespeichert bleiben, müssen sie nicht immer wieder neu eingegeben werden.

Beispiele:

- Die Umsetzung der DOS ASCII Umlaute in Windows ANSI Umlaute.
- Man sollte bestrebt sein, einheitliche Abkürzungen zu verwenden. Man findet z.B. als Abkürzung für das Wort ,,täglich“ sowohl ,,tägl.“ als auch ,,tgl.“. Man wird daher eine Ersetzungsanweisung eingeben, um alle Vorkommen von ,,tgl.“ in ,,tägl.“ umzusetzen.

NOFRAGMT

Ein Problem bei neuen Fachvokabularen ist leider, dass die gefundenen neuen Worte durch unerwünschte Wortfragmente verunreinigt sind, die durch die Silbentrennung am Zeilenende verursacht sind. NOFRAGMT erlaubt, diese Silbentrennungen weitgehend rückgängig zu machen. Es kann spezifiziert werden, in welchen Fällen dieses Rückgängigmachen nicht erfolgen soll.

DELLINES

Dieses Tool ist wichtig für die Anonymisierung der Quelltexte. Es arbeitet eng mit dem Tool TXTSTART zusammen. DELLINES erlaubt wahlweise:

- die ersten x Zeilen aller Dateien im Quellverzeichnis zu löschen,
- den Anfang aller Dateien im Quellverzeichnis bis zu einer bestimmten Textstelle zu löschen,
- ab einer bestimmten Textstelle den Text bis zum Ende derjeweiligen Datei zu löschen,
- den Text zwischen zwei Textstellen zu löschen,
- in allen Dateien des Quellverzeichnisses diejenigen Zeilen zu löschen, in denen ein bestimmter Text (z.B. ,,Frau“ oder ,,Herr“) vorkommt,
- alle Zeilen zu löschen, in denen ein bestimmter Text nicht vorkommt. Der Nutzen dieser Funktion wird im Zusammenhang mit dem Tool REMOVERR erläutert.

ABBREVNS

Dieses Tool erzeugt eine Liste allerAbkürzungen, die in den Dateien im Quellverzeichnis vorkommen. Diese Abkürzungen werden in die Datei ,,Abkürzungen.txt“ im Wortlistenverzeichnis eingetragen. In einem separaten Schritt ist dann die Liste der Abkürzungen manuell bzw. mit einem passenden TextPrep-Tool (2LISTS bzw. MODFYLST) halbautomatisiert mit der gesprochenen Form zu ergänzen.

INITIALS

Das Tool INITIALS erzeugt eine Wortliste aller Begriffe, die im Wort Großbuchstaben enthalten. Beispiele: ADAC, MHz, eMail. Diese Begriffe werden in die Datei ,,MehrGroßbuchstaben.txt“ im Wortlistenverzeichnis eingetragen. In einem separaten Schritt ist dann die Liste der Begriffe manuell bzw. mit einem passenden TextPrepTool (2LISTS bzw. MODFYLST) halbautomatisiert mit der gesprochenen Form zu ergänzen.

NOVOWELS

Mit diesem Tool werden alle Begriffe gefunden, die keine Vokale enthalten. Damit sollen möglichst viele Maßeinheiten gefunden werden. Beispiele: mg/ml, µg, °C, qm. Diese Begriffe werden in die Datei ,,VokalfreieBegriffe.txt“ im Wortlistenverzeichnis eingetragen. In einem separaten Schritt ist dann die Liste der Begriffe manuell bzw. mit einem passenden TextPrep-Tool (2LISTS bzw. MODFYLST) halbautomatisiert mit der gesprochenen Form zu ergänzen.

RENAME

Manche Quelldateien tragen in ihrem Namen Hinweise auf Personen. Zur Anonymisierung ist es daher wünschenswert, diese Dateien umzubenennen. Das Tool RENAME ändert alle Dateinamen des Quellverzeichnisses in einen vorgegebenen Namen, ergänzt mit einer fortlaufenden Zahl.

PHRASES

Um eine hohe Erkennungsrate zu erzielen, ist es für ein Fachvokabular sinnvoll, auch Begriffe wie z.B. ,,Hals- und Beinbruch“ oder ,,Knochenfraktur und -behandlung“ hinzuzufügen. Das Tool PHRASES extrahiert alle Ausdrücke dieser Art. Diese Begriffe werden in die Datei ,,Phrasen.txt“ im Wortlistenverzeichnis eingetragen. Sie werden auf Wunsch mit der gesprochenen Form automatisch ergänzt, damit diese Ausdrücke wie gewohnt ohne Ergänzungsstrich diktiert werden können.

MULTTERM

Sehr wichtig für Fachvokabulare sind Mehrwortbegriffe wie z.B. ,,Mutter-KindBeziehung“. Es soll dem Diktierenden erspart bleiben, die Bindestriche mitdiktieren zu müssen. Es ist daher sehr ratsam, diese Mehrwortbegriffe aufzufinden und in einer Wortliste mit der gesprochenen Form (ohne Bindestriche) bereitzustellen. MULTTERM findet diese Begriffe, ergänzt sie auf Wunsch automatisch mit der gesprochenen Form, und speichert sie in der Datei ,,Mehrwortbegriffe“ in das Wortlistenverzeichnis. Es kann festgelegt werden, ob auch Mehrwortbegriffe gefunden werden sollen, die ein anderes Sonderzeichen als der Bindestrich enthalten. So werden z.B. mit dem Zeichen ,,@“ alle E-Mail-Adressen gefunden.

COMBFILE

Dieses Tool fasst alle (TXT)-Dateien des Quellverzeichnisses zu 1 Datei oder einer frei wählbaren Anzahl von Dateien zusammen und legt sie im Zielverzeichnis unter dem Namen ,,GesamterText.txt“ ab.

REMOVERR

Die Liste neuer Wörter, die man z.B. mit Hilfe des Nuance VocTools oder einem der TextPrep-Tools erstellt hat, müssen in aller Regel überarbeitet werden Man korrigiert fehlerhafte Wörter, indem man mittels Auswahltasten die Schreibweise des Wortes ändert:

- nur der 1. Buchstabe ist groß geschrieben,
- alle Buchstaben sind groß geschrieben,
- das Wort wird klein geschrieben,
- die beiden Zeichen vor und nach dem Cursor werden getauscht,
- das Bindestrichwort wird zu 1 Wort zusammengefasst,
- das Bindestrichwort wird in 2 Worte getrennt,
- an der Stelle des Cursors wird das Wort in 2 Wörter mit Bindestrich getrennt mit wahlweise großgeschriebenem oder kleingeschriebenen 2. Wort
- Leerzeichen zwischen 2 Wörtern wird durch einen Bindestrich oder einem anderen, vorher spezifizierten Sonderzeichen ersetzt,
- das Wort(fragment) soll aus der Liste oder dem Quelltext entfernt (d.h. ersatzlos gelöscht) werden.
- Wahlweise wird der ausgewählte Begriff automatisch in die Windows Zwischenablage geschrieben. Dies erlaubt, wenn man das TextPrep-Tool CONTEXT parallel geöffnet hat, sofort die Textumgebung dieses Begriffes zu sehen.

Das markierte Wort befindet sich im Bearbeitungsfeld im Einfügemodus. Falls erforderlich, kann man es manuell beliebig verändern oder ergänzen.

Sind alle fehlerhaften Worte korrigiert, wählt man ,,Korrekturliste (auf Text) anwenden“, und Textprep überträgt automatisch alle Änderungen auf den ausgewählten Quelltext!

2LISTS

Bei den erzeugten Wortlisten Bindestrichbegriffe („Mutter-Kind-Beziehung“) und den Begriffen mit einem sog. Ergänzungsstrich („Auf- und Umbau“) wird die gesprochene Form auf Wunsch automatisch generiert.

Bei den Wortlisten Abkürzungen („Tabl.“), Maßeinheiten („mg/l“) und bei sog. Großbuchstabenworte („ADAC“) ist eine gesprochene Form unbedingt erforderlich, kann jedoch nur halbautomatisch erzeugt werden. Da diese Listen sehr lang sein können, ist es sehr zeitraubend, den einzelnen Wörtern die gesprochene Form manuell hinzuzufügen.

Es kommt noch hinzu, dass beim Erstellen von Fachvokabularen mehr als 80 Prozent dieser Begriffe und ihre gesprochene Form immer wieder vorkommen. Das Tool 2LISTS erlaubt nun, bei einer neu erstellten Liste die gesprochene Form aus früheren Listen auf leichte und sehr zeitsparende Weise zu übernehmen.

NEXTWORD

Adjektive und Verben werden oft im Zusammenhang mit den Wörtern "alles, etwas, nichts, ...usw. als Substantive verwendet. Diese werden jedoch bei Dragon NaturallySpeaking oft klein geschrieben. Dies kann verhindert werden, wenn man diese Wörter als Begriff in das Vokabular aufnimmt, so z.B. "alles Schöne, nichts Gutes, etwas Erfreuliches,...usw.).

Umgekehrt möchte man auch Mehrwort-Ausdrücke wie z.B. "Diabetes mellitus" erfassen. NEXTWORD findet diese Begriffe ebenfalls. Alle diese Begriffe können dann dem Vokabular als eigenständige Ausdrücke hinzugefügt werden.

XPRESSNS

Es gibt im Deutschen oft Namen von Organisationen usw., die aus mehreren großgeschriebenen Wörtern bestehen, ggf. auch mit kleingeschriebenen Füllwörtern. Beispiele: "Deutsches Rotes Kreuz", "Internationales Institut für Weltraumforschung". XPRESSNS erzeugt aus den Quelltexten eine Wortliste solcher Ausdrücke.

MODFYLST

Mittels dieses Tools kann man in einer Wortliste

- mittels MS Word und/oder einer anderen Rechtschreibprüfung die Wörter teilen in "richtig geschrieben" und "vermutlich falsch geschrieben". "Vermutlich falsch geschrieben" heißt nur, das dieses Wort der Rechtschreibprüfung unbekannt ist;
- die gesprochene Form erzeugen und/oder selektiv die gesprochene oder die geschriebene Form bearbeiten;
- auch bei z.B. englischen oder französischen Fremdwörtern die gesprochene Form erzeugen;
- ausgewählte Wörter und deren gesprochene Form in eine externe Datei auslagern;
- Textstrings am Anfang, am Ende oder irgendwo im Wort suchen und global ersetzen; sowohl in der geschriebenen oder in der gesprochenen Form;
- selektierte Wörter löschen;
- die gesprochene Form von Maßeinheiten automatisch generieren.
  Beispiel: mg/cm2 mg/cm2\Milligramm pro Quadratzentimeter

SPACED

Dieses Tool listet in einer Datei alle Begriffe, die im Quelltext in gesperrt geschriebener Form vorkommen. Dies erlaubt mittels des TextPrep-Tools REMOVERR diese Begriffe in ihre normale Schreibweise zurückzuführen, um sie so den Wortanalysen bzw. -listen zugänglich zu machen.