



Vorstellung des PetaMem NLP Portals

Richard Jelinek
PetaMem, s.r.o.
Ocelářská 1, Prag, Tschechien
rj@petamem.com

Angefertigt am: 5.1.2005

Abstract

Es werden Konzepte, Funktionalität und technologische Grundlagen des PetaMem NLP/NLU Portals vorgestellt. NLP steht für “natural language processing” und umfasst alle Verfahren zur Verarbeitung natürlicher Sprache, NLU steht für “natural language understanding” und umfasst Verfahren und Methoden, welche das semantische Verständnis natürlicher Sprache durch Computer zum Ziel haben.

Das Portal ist erst in seinen Anfängen und der sichtbare Teil zeigt bislang nur eine Teilfunktionalität des Zusammenspiels dreier Systemkomponenten mit einer Gesamtkomplexität von knapp 650.000 LOC (ohne externe CPAN Module).

Schlüsselwörter:

Perl, Mason, Portal, Sprachverarbeitung, Sprachverstehen, maschinelle Übersetzung, Sprachidentifikation

PetaMem Copyright Notice

Copyright © 2002-2005 PetaMem, s.r.o. Alle Rechte vorbehalten.

Dieses Werk ist geistiges Eigentum der PetaMem, s.r.o. Es darf als Ganzes oder in Auszügen kopiert werden, vorausgesetzt, daß sich dieser Copyright-Vermerk auf jeder Kopie befindet.

1 Einleitung

Von einer vagen Idee Anfang 2003 “Website mit dynamischem NLP Inhalt” hat sich das hier besprochene NLP Portal 2005 zu einem zentralen und strategisch wichtigen Projekt gemauert. Drei gewichtige Gründe stehen hinter diesem Projekt:

Webclient für Corporate Intranet Unser Kernprodukt - PMLS¹ findet vor allem in großen mittelständischen sowie in Großunternehmen Verwendung. Die Nutzung erfolgt vorwiegend durch Mailclients sowie ein Webfrontend. Der verstärkte Wunsch der Kunden eine Möglichkeit zur Integration im firmeneigenen Intranet zu bieten verlangte zunehmend die Bereitstellung eines generischen Webfrontends anstelle vieler individueller und proprietärer “Einzelanfertigungen”.

Demonstrationsplattform Die Kosten für ein PMLS System sind im höheren 5-stelligen EUR-Bereich angesiedelt. Selbstverständlich geht einer Kaufentscheidung eine - recht arbeitsintensive - Beratung und Demonstration voraus. Der Verweis auf die Portalfunktionalität reduziert diesen Beratungsaufwand erheblich. Zugriff auf spezifische Funktionalität für Benutzerklassen erlaubt überdies auch individuelle Demos.

komplementäre Einnahmequelle Die Software ist für kleinere Unternehmen - geschweige denn für Privatpersonen - nicht erschwinglich bzw. nicht wirtschaftlich. Alleine die Hardwareanforderungen stellen einen beträchtlichen Kostenfaktor dar. Die Leihstellung von PMLS Systemen eliminiert zwar die Investitionskosten, erfordert für ihre Wirtschaftlichkeit jedoch ein entsprechendes Nutzungsvolumen. Ein Pay-per-use revenue Modell erlaubt den Zugriff auf diese Funktionalität schon ab geringen Centbeträgen.

Darüber hinaus erscheint es realistisch das Portal zu einer “Prestige - Demonstrationsplattform” ausbauen zu können, anhand derer nicht nur Kunden oder konkret Interessierte Funktionalität testen können, sondern allgemein Interessierten eine Technologiekompetenz vorgeführt werden kann, welche sich positiv auf das Firmen-Image auswirken könnte.

2 Technische Aspekte des Portals

2.1 Geschichtliches

Der ursprüngliche Gedanke “eine Website mit dynamischem NLP Inhalt”, sah keinesfalls ein derart umfangreiches System vor. Daher wurde nach einem Vehikel gesucht, welches eine entsprechende Funktionalität mit möglichst geringem Aufwand realisieren könnte. Die Wahl fiel auf Yawps², aufgrund einiger Zusagen in der Projektbeschreibung (keine DB notwendig, modperl - gar modperl2 kompatibel). Die karge NLP Funktionalität wurde in neu erstellte (Cut&Paste) Yawps Module hardcodiert.

Bis Ende 2003 erfuhr diese Lösung zahlreiche Erweiterungen und wurde “aufgebohrt”, indem eine Schnittstelle zu MySQL geschaffen wurde (vom Projekt explizit nicht vorgesehen) um das Usermanagement hierauf aufzubauen. Des weiteren wurde eine Synchronisation mit unserem CRM implementiert.

Dennoch wurde zunehmend offensichtlich, daß Yawps nicht über die benötigte Infrastruktur und Mächtigkeit verfügt um die stetig wachsenden Ansprüche

¹PetaMem Language Server

²<http://yawps.sourceforge.net>

an diese Webpräsenz erfüllen zu können. Im Mai 2004 - aufgrund sich abzeichnender Anforderungen an das Portal - dann die Entscheidung einer Reimplementierung auf Basis von Perl/Mason. Einen detaillierten Überblick liefert auch http://nlp.petamem.com/help.cgi?tid=15&new_site_lang=de&inline=1 (<http://www.tinyurl.com/53uay>)

2.2 Zentrale Komponenten

Für die Funktion des Portals werden drei Komponenten benötigt:

1. Die Portalsoftware selbst (Perl/Mason, modperl2, Apache2)
2. PMLS Server als Backend für NLP Funktionalität (Perl)
3. CRM/ERP für User Management, Preislisten und als Marketing-Tool (Perl)

Diese Komponenten haben eine Gesamtkomplexität von knapp 650.000 LOC³ - bzw ein Äquivalent von 20 Mannjahren Entwicklungsaufwand⁴. Bild 1 zeigt grob das Zusammenspiel dieser Komponenten. Hierbei erfolgt eine reguläre HTTP Anfrage vom User/Browser/Crawler, NLP baut die entsprechende Seite auf. Ist hierzu die Anfrage an den PMLS Server erforderlich (weil die Darstellung bereits das Ergebnis einer NLP Funktionalität erfordert), agiert NLP als Client eines PMLS Servers, stellt seine Anfrage und erhält die Antwort welche er dann entsprechend aufbereitet darstellt. Die Kommunikation zwischen NLP und CRM ist gestrichelt angezeigt, da diese nur manuell initiiert wird um dann mittels Push/Pull die Daten en Block abzugleichen.

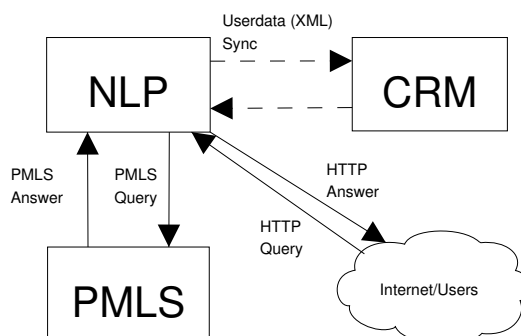


Figure 1: Zusammenspiel von NLP, PMLS und CRM.

Beim NLP/CRM Abgleich werden nicht nur die Benutzerdaten, sondern auch etwaige Preislisten, sowie mit der Benutzerverwaltung zusammenhängende Hilfsdaten aktualisiert (z.B. Länder liste, Preislisten, Firmentypen in versch. Ländern u.a.).

2.3 Verteilte Client/Server Architektur

PMLS selbst verfügt über eine C/S Architektur, bei welcher der Server ein `Net::Server::Multiplex` Derivat ist und entsprechend mehrere Clients im Multiplexverfahren bedient. Darüber hinaus können auf einer Maschine "beliebige viele" (natürlich im Rahmen vorhandener Systemressourcen) Serverinstanzen gestartet werden.

³Stand 30.12.2004 - NLP 125', PMLS 400', CRM 123'

⁴Die CRM/ERP Komponente wurde jedoch von PetaMem eingekauft und an die eigenen Bedürfnisse angepasst

Da die Clients selbst über einen Load-Balancing Mechanismus verfügen mit dem zwischen verschiedenen lokalen oder remote Servern gewählt werden kann, ergibt sich ein äußerst flexibles verteiltes C/S Konzept.

Ein Server kann sich zwar temporär forken (Parallel::ForkManager), jedoch nicht um so parallel mehrere Clients zu bedienen, sondern um ggf. leicht parallelisierbare Aufgaben für einen Client schneller auszuführen. Die Software ist für die Ausführung auf einem eng gekoppelten System konzipiert (UP oder SMP) und verfügt über keinen eigenen Support für Cluster. Der Grund hierfür ist die Bereitstellung einer entsprechenden Infrastruktur durch Mosix⁵.

2.3.1 Online und Offline Bezahlendienste

Die Einbindung von Online- sowie Offline Bezahlendiensten ist für jedes Portal, welches auch kommerzielle Dienste anbieten möchte von zentraler Bedeutung. Die heutige Problematik besteht im Wesentlichen darin, daß jeder Bezahlendienst nur einen Teil der potentiellen (und letztlich auch realen) Nutzer anspricht, gleichzeitig jedoch Kosten aufweist, welche nur wirtschaftlich erscheinen, wenn ein bestimmtes Mindest-Nutzungsvolumen erreicht wird.

Darüber hinaus sind die verfügbaren Bezahlmethoden jenseits von Kreditkarte & Co. stark länderspezifisch und können auf einem Portal mit internationaler bzw. globaler Ausrichtung nicht grundsätzlich angeboten werden.

Wir haben uns daher entschieden ein generisches Payment-Framework zu implementieren, bei dem die spezifischen Bezahlmethoden je nach Verfügbarkeit bzw. Freischaltung angeboten werden. Dies wird aufgrund des Wohn- sowie Aufenthaltsortes des Benutzers ermittelt. So erhalten beispielsweise Benutzer (ab Registered) mit Wohnort Tschechien, als mögliche Bezahlmethoden lokale Dienste (online eBanka - <http://www.ebanka.cz> bzw. offline Kontonummer zwecks Überweisung) sowie globale Dienste welche auch für CZ verfügbar sind (z.B. PayPal - <http://www.paypal.com>). Benutzer aus Deutschland hingegen erhalten neben den global verfügbaren Diensten wieder die Möglichkeit z.B. mittels T-Pay oder Firstgate zu bezahlen⁶.

Um den internationalen Benutzern einen Überblick der - intern in EUR gehaltenen - Kosten für die angebotenen Dienste zu ermöglichen, werden tagesaktuelle Umrechnungskurse von der EZB geholt⁷ und dem Benutzer alternativ zu den EUR Kosten die entsprechenden Beträge auch in "seiner" Währung angezeigt.

2.3.2 Reichweite: Browser, Themen, Lokalisierung

Das Portal hat schon durch seine Thematik implizit eine internationale Ausrichtung. Um dieser gerecht zu werden muss selbstverständlich einem möglichst breiten Benutzerkreis ein reibungsloser Zugang ermöglicht werden. Hierzu gehört eine größtmögliche Browserkompatibilität, die Bereitstellung von Themen (Look&Feel) für bekanntlich verschiedene Geschmäcker sowie die vollständige Lokalisierung und Internationalisierung sowohl des Portals selbst, wie auch der Portal-Inhalte.

Es ist uns kein weiteres Portal bekannt, welches all diese Anforderungen in einem gemeinsamen Framework zu erfüllen versucht. Bei der Implementierung wurde auch schnell klar warum. Die sich hieraus ergebende Kombinatorik⁸ macht diese Aufgabe nicht gerade trivial und erfordert zu ihrer Realisierung ein mächtiges Framework.

⁵<http://www.mosix.org/> bzw. <http://openmosix.sourceforge.net/>

⁶Gegenwärtig sind diese Bezahlmöglichkeiten wegen der geringen Akzeptanz deaktiviert

⁷siehe XML Feed auf <http://www.ecb.int/stats/eurofxref/eurofxref-daily.xml>

⁸Anzahl der unterstützten Browser x Anz. Themen x Anz. Sprachen - hinzu kommt jedoch in einigen Fällen noch als Faktor die Anzahl der Benutzerklassen, da sich z.B. Eingabeformulare je nach Benutzerklasse unterscheiden können.

2.3.3 Steuerliche Aspekte

Nun ist die internationale Bereitstellung von Bezahldiensten durch ein international tätiges Unternehmen so lange keine echte technische Herausforderung, bis sich ein oder mehrere nationale Finanzämter zu Wort melden. Ein nicht-trivialer Mix aus folgenden Parametern entscheidet darüber ob und falls ja in welcher Höhe Mehrwertsteuer auf die angebotenen Dienste erhoben wird und wo das Einkommen des Unternehmens aus dieser Tätigkeit zu versteuern ist:

- Sitz des Unternehmens
- Aufstellungsort des/der Server(s)⁹
- Vom Benutzer angegebener ständiger Wohnort
- Aufenthaltsort des Benutzers zum Zeitpunkt der Inanspruchnahme der Onlinedienste

Alle Informationen sind verfügbar, wobei bei dem angegebenen Wohnort des Benutzers auf Plausibilitätsprüfungen sowie die Korrektheit der Selbstauskunft vertraut werden muss, genauso ist der ermittelte Aufenthaltsort des Benutzers zum Zeitpunkt der Anfrage mittels GeoIP festgestellt und kann nicht garantieren, daß es sich auch tatsächlich um den physischen Aufenthaltsort des Benutzers handelt (Anonymisierungsdienste, Corporate-Gateway eines transnationalen Unternehmens,...), für den regulären Betrieb, wird jedoch die von GeoIP gelieferte Lokation als korrekt angenommen.

2.4 Hard- und Software Infrastruktur

Bild 2 zeigt die Anfangskonfiguration des Portals mit jeweils einem NLP und PMLS Server und einigen wenigen global verteilten Picservern. Die Picserver liefern ausschließlich statischen Content und zwar je nach geographischem Aufenthaltsort des zugreifenden Benutzers (festgestellt mittels GeoIP).

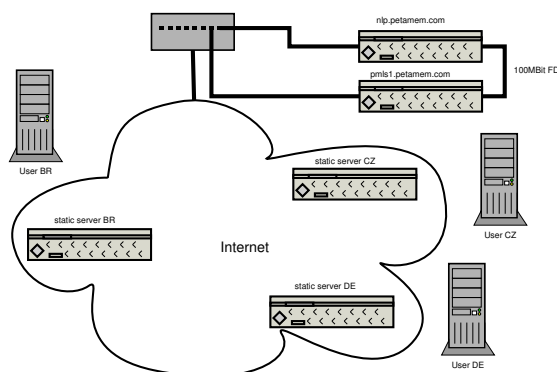


Figure 2: Schema der gegenwärtigen verteilten Konfiguration.

Sowohl um die Antwortzeiten zu minimieren wie auch die Traffic-Last vom zentralen NLP Server zu nehmen bzw. zu dezentralisieren, enthält NLP eine Komponente um z.B. Verweise auf Grafiken je nach Aufenthaltsort des Benutzers an externe Server zu delegieren. Hierbei werden einfach Schablonen mit den URLs kurz vor Auslieferung mit der entsprechenden URL belegt. Picserver können vom NLP Admin zur Laufzeit hinzugenommen oder entfernt werden. Darüber hinaus prüft das Portal ihre Erreichbarkeit (siehe Bild 3).

⁹Wer es deftig mag, darf gerne eine Menge von Aufstellungsorten eines dezentralen Systems in Betracht ziehen



Figure 3: Picserver hinzunehmen/verändern/löschen.

Bild 4 zeigt eine hoch skalierte Konfiguration, mit Lastverteilung bei Picservern und sowohl bei den NLP wie auch bei den PMLS Servern (vgl. auch Abschnitt 2.3) und dediziertem DB Server, wie sie an einem Aufstellungsort möglich ist. Die Leistung dieser Konfiguration ist abhängig von der Anbindung der NLP Server ans Internet. Sollte sich irgendwann auch diese Konfiguration als Flaschenhals herausstellen, ermöglicht die verteilte Systemarchitektur die Nutzung des DistributedDNS von Akamai, erfordert dann jedoch ebenfalls dezentrale und replizierende DB Server.

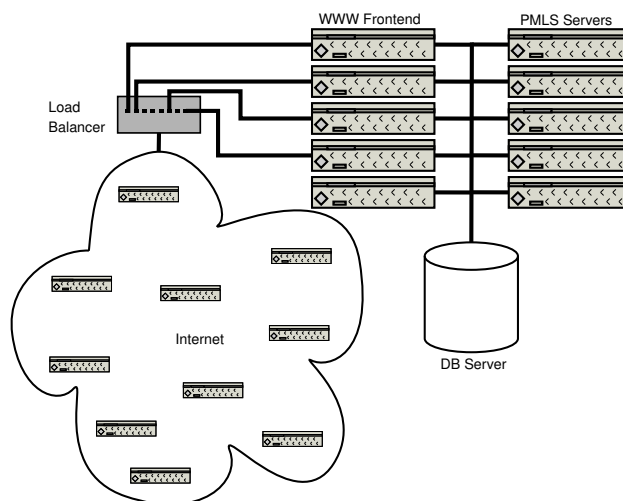


Figure 4: Schema einer skalierten verteilten Konfiguration.

3 NLP/NLU Funktionalität

Hauptaufgabe des NLP Portals ist es, NLP und NLU Funktionalität online zur Verfügung zu stellen, indem ein Frontend für verschiedenen NLP/NLU Dienste (seien diese von PetaMem oder von Drittanbietern) angeboten wird. Diese Dienste unterscheiden sich hinsichtlich Komplexität und damit verbunden in ihrem Ressourcenbedarf. Im Wesentlichen gibt es **Basisdienste** sowie darauf aufbauend komplexere **wortbasierte** und **textbasierte** Dienste. Die unter einer “Fun-Sektion” angebotene Funktionalität stellt gewissermaßen ein Nebenprodukt dar.

3.1 Basisdienste

Basisdienste stellen - wie der Name schon verrät - grundlegende Funktionalität bereit, welche in vielen NLP/NLU Anwendungen Grundvoraussetzung für die Bereitstellung komplexerer Dienste ist. Hierzu gehören insbesondere Sprachidentifikation, Textrekonstruktion u.a.

3.1.1 Sprachidentifikation

Die automatische Erkennung der Sprache eines vorgegebenen Textes ist eine der grundlegenden, zugleich jedoch wichtigsten Eigenschaften sprachverarbeitender Systeme. Überdies eröffnet dieses Feature für sich alleine genommen ein breites Spektrum möglicher Anwendungen. Bild 5 zeigt die Eingabemaske dieses Moduls.

Sprachidentifikation Hilfe ?

Sprachidentifikation

Erkennungsmethode: NGram

Details der Sprachidentifikation anzeigen

Text eingeben oder mittels cut&paste einfügen:

You have successfully registered to the German Perl-Workshop 7.0. Please transfer the fee within one week and don't forget to mention your registration number (see below).

oder wählen Sie ein File-Upload hier. Browse...

Servicekosten: 29 Einheiten (0.029 EUR)
Verfügbare freie Einheiten: 150

Reset
OK

Figure 5: Identifikation der Sprache eines Textes.

Der zu bearbeitende Text kann entweder in das entsprechende Textfeld eingeben (oder mittels Cut&Paste kopiert) werden, oder als Textdatei vom lokalen PC des Benutzers hochgeladen werden. Für den Fall, daß sowohl ein Text eingegeben wie auch eine Datei zum hochladen ausgewählt wird, wird der eingegebene Text verworfen und nur die Datei bearbeitet.

Die bereitgestellten Texte werden vollautomatisch verarbeitet und nach der Analyse verworfen. Diese Texte werden weder gespeichert, noch für andere Zwecke verwendet.

Der Identifikationsprozess kann auch durch Auswahl der "Erkennungsmethode" gesteuert werden. Diese kann einen der folgenden Werte annehmen:

Dict Eine wörterbuchbasierte Methode der Sprachidentifikation. Der gegebene Text wird wortweise iteriert und das System versucht die gefundenen Wörter in den Systemwörterbüchern zu finden. Diese Methode ist recht rechenaufwändig, jedoch gut geeignet für kurze Texte, wo statistische Methoden oft fehlschlagen.

NGram Dies ist der Industriestandard zur Sprachidentifikation. NGram-basierte Sprachidentifikation ist schnell und zuverlässig (falls man über gute so genannte "Sprachmodelle" verfügt) für mittellange Texte.

NVect Dies ist eine proprietäre Technologie von PetaMem zur Identifikation der Sprache eines gegebenen Textes. Es handelt sich dabei um eine Generalisierung der NGram Methode, welche einen erhöhten Rechenzeitbedarf aufweist und nicht für lange Texte geeignet ist, jedoch bessere Ergebnisse als NGram bei kürzeren Texten liefert.

Smart Wie schon der Name dieser standardmäßig aktivierten Methode sagt, wird hierbei versucht zu derjenigen Erkennungsmethode zu wechseln, welche am geeignetsten für einen gegebenen Text ist. Dies sollte die beste Wahl für alle Fälle sein.

Wird die "Zeige Details" checkbox aktiviert, wird das Ergebnis detaillierter beschrieben. Die Informationen hängen von der gewählten Methode ab.

Der Erfolg der Sprachidentifikation hängt von der "Reinheit" des eingegebenen Textes ab. Das Ergebnis der Identifikation könnte abweichen, falls Text eingegeben wird, welcher Markup enthält (z.B. HTML, LaTeX, ...).

3.1.2 Diakritik Prozessor/Textrekonstruktion

Bild 6 zeigt die Eingabemaske des Moduls zur Rekonstruktion oder Entfernen der Diakritik eines Textes. Diese Funktionalität ist interessant um Texte von oder in Formate zu überführen, welche keine vollständige Darstellung des Zeichensatzes erlauben. Z.B. deutsche Mails ohne Umlaute, tschechische SMS ohne Diakritik etc. Es ist also nur von Nutzen bei Sprachen, welche über eine Diakritik verfügen, wie z.B. Deutsch, Tschechisch und andere. Englisch und Latein z.B. verfügen über keine Diakritik.

Diakritik

Diakritik-Modus: Choose

Hinweis Textsprache: Tschechisch

Text eingeben oder mittels cut&paste einfügen:

Korpus je soubor počítačové uložených textů (v případě mluveného jazyka - prepisu záznamu mluvy), který slouží k jazykovému výzkumu. K práci s tímto korpusem slouží speciální vyhledávací program. S jeho pomocí je možné vyhledávat slova a slovní spojení v kontextu a zjistit jejich frekvenci v korpusu i původní textový zdroj. Umožňuje i další zpracování nalezeného (např. abecední třídění apod.). U některých korpusů lze vyhledávat i podle slovních druhů.

oder wählen Sie ein File-Upload hier. Browse...

Servicekosten: 201 Einheiten (1.005 EUR)
Verfügbare freie Einheiten: 150 [Diakritik], 150 [Sprachidentifikation]

Reset

OK

Figure 6: Rekonstruktion oder Entfernen der Diakritik.

Texthandling entsprechend dem Sprachidentifikation-Modul (vgl. Abschnitt 3.1.1). Das Modul kennt 3 Betriebsarten:

1. **Fit1st** Eine Art der Rekonstruktion. Die wahrscheinlichste Wortalternative (ob nun mit oder ohne Diakritik) wird benutzt. Da die "wahrscheinlichste" Alternative (siehe Einschränkungen weiter unten) nicht immer die Richtige ist, gibt es:
2. **Choose** Eine Art der Rekonstruktion. Falls es mehr als eine Alternative gibt, werden dem Benutzer alle angeboten, damit er seine Wahl trifft. Obwohl dies die meiste Interaktion auf Benutzerseite erfordert, entspricht die angebotene Voreinstellung der Option 'Fit1st' und ist in mehr als 90% aller Fälle korrekt.
3. **Remove** Entfernen der Diakritik. Jegliche Diakritik wird von einem gegebenen Text entfernt. Dies ist die inverse Operation zur Rekonstruktion.

Bild 6 zeigt die angebotene Maske nach erfolgreichem Durchlauf der Rekonstruktion eines tschechischen Textes an bei Wahl der Option "Choose". Wäre Option "1stFit" gewählt worden, wäre der Text so wie er zu sehen ist, ohne weitere Wahlmöglichkeit fertig für Cut&Paste bzw. Download angeboten worden.



Figure 7: Auswahl der Alternativen bei Option "Choose".

Für die Rekonstruktion der Diakritik muss das System die Sprache des betreffenden Textes kennen. Per Default versucht das System die Sprache des Textes automatisch zu identifizieren. Durch das Setzen des Hinweises für die Textsprache, werden sowohl Systemressourcen (und folglich Einheiten) gespart und gleichermaßen mögliche Unsicherheiten des Identifikationsprozesses eliminiert.

3.1.3 Sprachspezifische Zeichensätze

Dieses kleine Java Script Popup (siehe Bild 8) stellt eine nützliche Hilfsfunktion dar um bei Bedarf die passenden Zeichen zur Verfügung zu haben, sofern die Tastatur diese nicht hergibt. Gegenwärtig werden die Zeichensätze von 37 Sprachen unterstützt.

3.2 Fun

Die generische Funktionalität eines PMLS Servers im Bereich Sprachverarbeitung/Sprachverstehen ermöglicht eine Vielzahl von Anwendungen aus dem Bereich Spaß&Spiel bzw. Zeitvertreib. Hierzu zählen - allesamt multilinguale - Anwendungen wie:



Figure 8: Popup für sprachspezifische Zeichensätze.

- Anagramm Server (auch Cross-Lingual)
- Palindromsuche
- Generische Wortsuche (mehrere Ausprägungen wie z.B. Scrabble, Kreuzworträtsel, Reime u.a.)
- Edutainment (Vokabeltrainer, Quiz, etc.)

Als Termin für die Realisierung eines individuellen Webfrontends dieser Funktionalitäten wurde “wenn mal Zeit ist” anberaumt. Dieser PetaMem-interne Running Gag bedeutet eher mittel- und noch eher langfristige Realisierung. Dennoch ist z.B. die generische Wortsuche für erfahrene Benutzer anderweitig bereits zugänglich - nämlich in der Regex Suche des M³ Wörterbuches.

3.3 Wortbasiert

3.3.1 Prozedurale Wörter - Number ↔ Word

Das “Number ↔ Word” Modul, dessen Frontend in Bild 9 dargestellt ist, soll die Verarbeitung von so genannten prozeduralen Wörtern veranschaulichen. Hierbei handelt es sich um solche Wörter¹⁰, die algorithmisch gebildet werden und weder mittels Morphologie beschrieben, noch in einem Lexikon bzw. Wörterbuch aufgelistet werden können. Ein derartiges Beispiel sind in allen Sprachen Numerale.

Die Aufgabenstellung eine Dezimalzahl in ein Numeral umzuwandeln ist relativ einfach und z.B. von der “in Worten”-Darstellung eines Betrages auf Schecks bekannt. Auch gibt es hierfür auf CPAN für verschiedene Sprachen Module¹¹ - 17 von diesen werden vom Frontend auf der Ausgabeseite (Zahl→Numeral) unterstützt.

Die Analyse eines Numerals und Konversion in eine Dezimalzahl ist schon schwieriger und erfordert den Einsatz eines Parsers für diese Wortklasse. Das Frontend unterstützt gegenwärtig mindestens 10 Sprachen auf der Eingabeseite (Numeral→Zahl) - sowie automatische Erkennung.

Der Hauptvorteil dieses Ansatzes ist, daß durch die Umwandlung in Dezimalzahlen eine erstklassige Interlingua zur Verfügung steht und bei etwaigen Übersetzungen die Anzahl der

¹⁰bzw. allgemein POS - “part of speech”

¹¹vgl. `Lingua::Num2Word.pm` welches als Container-Modul ein einheitliches API zu diesen Modulen bietet

Nummer <=> Text Hilfe ?

Konversion der numerischen und textuellen Repräsentation von Numeralen

Eingabe: Deutsch

Ausgabe: Chinesisch

Servicekosten: 2 Einheiten (0.010 EUR) Dozent

Figure 9: Number ↔ Word Eingabeformular.

benötigten Transformer von $O(n^2)$ auf $O(n)$ sinkt. Dadurch werden in der gegenwärtigen Konfiguration über 150 Sprachpaare unterstützt.

3.3.2 M³ Wörterbuch

Der Name des in Bild 10 abgebildeten M³ Wörterbuchs ist von den drei M's abgeleitet:

1. Multifunktional - es ermöglicht verschiedene Suchverfahren: Exakt, Teil-von, Fuzzy (ähnlich), Regex (eingeschränkte reguläre Ausdrücke)
2. Multilingual - erlaubt die Suche in vielen verschiedenen Quell- und Zielsprachen
3. Morphologisch - erlaubt die Suche auch bei Eingabe von flektierten oder abgeleiteten Wortformen.

Das Wörterbuch auf dem NLP Portal weist die folgenden technischen Daten auf:

- unterstützt gegenwärtig über 80 Sprachpaare
- Sprachunterstützung zur Laufzeit erweiterbar
- unterstützt mehr Abfragearten als jedes andere Online Wörterbuch
- die Wörterbücher enthalten weit mehr als 1 Million Einträge
- Übersetzung von Phrasen möglich

3.4 Textbasiert

Die vom PMLS Server unterstützten NLP Operationen auf Texten sind sehr zahlreich und eine vollständige Diskussion würde den Rahmen dieses Dokumentes bei weitem sprengen. Exemplarisch seien daher zwei der häufigsten bzw. pragmatischsten Funktionen aufgeführt: Rechtschreibkorrektur sowie die "Königsdiziplin" Maschinelle Übersetzung.

Andere Funktionalitäten sind z.B.:

Figure 10: M³ Wörterbuch Eingabeformular.

Textkategorisierung Moderne statistische Verfahren ermöglichen das Training von Systemen, bei dem eine Menge von Texten manuell zu den entsprechenden (beliebigen) Kategorien zugeordnet wird. Anschließend führt das System eine Zuordnung von unbekanntem Texten diesen Kategorien vollautomatisch zu. Zusammen mit der Sprachidentifikation (vgl. Abschnitt 3.1.1) können so beispielsweise sehr leistungsfähige Dispatcher erstellt werden, welche vollautomatisch eingehende Mails korrekten Empfängern zuordnen. (siehe z.B. auch: <http://www.otrs.de/produkte/nlp/>)

Text Summarization Bei einem vorherrschenden Überangebot an Informationen, ist es sehr nützlich, wenn man von einem langen Text eine Zusammenfassung des Inhaltes bzw. der relevanten Fakten erhalten könnte. Also z.B. eine 2-seitige Zusammenfassung eines 50-seitigen Textes.

Sprach-/Wissensacquire Die Sprachidentifikation als statistische Methode kann mit einem vorhandenen Korpus einer vormals unbekanntem Sprache ein sog. Sprachmodell erzeugen und so neue Sprachen identifizieren. Dies ist oftmals nützlich um eine hohe Differenzierung bei verwandten oder ähnlichen Sprachen zu erreichen, oder um einfach seltene, jedoch im eigenen Alltag wichtige, Sprachen identifizieren zu können.

Q/A System/Chatbots Die vorgenannten Verfahren konzentrieren sich insbesondere auf die Analyseseite der Sprachverarbeitung. Ebenso wichtig ist jedoch die Generierung natürlicher Sprache, wie sie z.B. bei der maschinellen Übersetzung von Texten, aber auch in Dialogsystemen benötigt wird. Intelligente Chatbots, die z.B. auf eCommerce Webseiten als Berater fungieren können und auf gestellte Fragen *intelligente* Antworten liefern können, sind ein klassischer Anwendungsfall.

3.4.1 Rechtschreibprüfung

Eine Rechtschreibprüfung gehört heutzutage zur Standardausstattung jedes Office Paketes und auch jeder bessere Freemailer bietet einen Spellchecker.

Alle gegenwärtigen Systeme gehen jedoch nicht weit über eine Rechtschreibkorrektur im Wortkontext hinaus. So werden z.B. Sätze wie “This is the last think to do.” als korrekt erkannt, da natürlich “think” ein korrektes Wort ist und im Wortkontext kein Fehler er-

sichtlich ist. Bindet hingegen eine Rechtschreibkorrektur eine syntaktische Satzanalyse mit ein wird der o.g. Fehler entdeckt, da bereits die Wortklasse (Verb statt Nomen) nicht stimmt.

Eine weitere Verbesserung der Güte der Rechtschreibprüfung erhält man durch die statistische Analyse großer Korpora, welche die Wahrscheinlichkeit von Kollokationen - also des gemeinsamen Auftretens von Wortkombinationen innerhalb von Phrasen - bestimmen. Somit können Aussagen darüber getroffen werden, ob "vermutlich nicht" oder "vermutlich Nacht" wahrscheinlicher ist, falls der Benutzer im Text "vermutlich nacht" eingegeben hat.

Weitere Verfahren wie z.B. tiefensemantische Analyse des zu überprüfenden Textes steigern die Erkennungsrate von Fehlern abermals¹². Alle diese Maßnahmen führen jedoch dazu, daß bei dem gegenwärtigen Stand der Technik diese Funktionalität nicht auf Seiten des Clients vorgehalten werden kann und entsprechend serverseitig auf einer ausreichend dimensionierten Maschine erfolgen muss. Bild 11 zeigt die Schnittstelle der Rechtschreibprüfung des NLP Portals.

Seite unbekannt Hilfe ?

Statistische Rechtschreibprüfung

*Language:

*Mode of operation:

Text eingeben oder mittels cut&paste einfügen:

С первых страниц газет и журналов всего мира так до сих пор и не сходят фотографии и сообщения из юго-восточного региона Азии. Ужасающая катастрофа отодвинуло на второй план все, что совсем недавно считалось самым важным и актуальным. Мощное цунами, спровоцированное девятибалльным поддонным землетрясением в Индийском океане, обрушилось на 12 стран региона, убило больше 150 тысяч человек, не считая пропавших без вести, нанесло многомиллиардный материальный ущерб, а теперь грозит массовыми эпидемическими заболеваниями всему человечеству

oder wählen Sie ein File-Upload hier.

Servicekosten: UNITS (MONEY)
Verfügbare freie Einheiten:

Figure 11: Rechtschreibprüfung.

Der Benutzer kann nun die Sprache auswählen - oder automatisch erkennen lassen, sowie einen von zwei Verarbeitungsmodi auswählen:

Mark hierbei werden die als fehlerhaft erkannten/vermuteten Wörter farblich markiert und es vollständig dem Benutzer überlassen mittels eines Texteditors die Fehler zu korrigieren.

Choose hierbei werden - ähnlich wie bei den Diakritik Operationen (s.o.) die als fehlerhaft erkannten/vermuteten Wörter zusammen mit den von der Rechtschreibprüfung vorgeschlagenen Alternativen in einem Drop Down Menü zusammengefasst und dem Benutzer zur Auswahl vorgelegt.

¹²Erhöhen aber zugegebenermaßen auch die Anzahl der Fehlalarme

3.4.2 Textübersetzung

Bild 12 zeigt das Modul für Textübersetzungen. Es stellt prinzipiell ein B2B bzw. B2C Portal bereit, bei dem Übersetzungsleistungen schnellstmöglich in gewünschter Qualität zu einem entsprechenden Preis ausgeführt werden.

Seite unbekannt Hilfe ?

Textübersetzung

Maschinelle Übersetzung *MT Quality: Simple
 Editierte maschinelle Übersetzung Zielsprache: Czech
 Menschliche Übersetzung

Text eingeben oder mittels cut&paste einfügen:

oder wählen Sie ein File-Upload hier.

Servicekosten: UNITS (MONEY)
Verfügbare freie Einheiten:

Figure 12: Eingabeformular für maschinelle und menschliche Übersetzung.

Benutzer können zwischen 3 Hauptmethoden der Übersetzung wählen:

Maschinelle Übersetzung Diese Option bewirkt, daß der eingegebene Text von einem Computer übersetzt wird. Abhängig von der zugrunde liegenden Software, können auch verschiedene Qualitätsstufen der Übersetzung (s.u.) gewählt werden.

Handkorrigierte Maschinelle Übersetzung Dies stellt einen sehr guten Kompromiss zwischen Kosten und Qualität dar. Der eingegebene Text wird von einem MÜ-System wie oben beschrieben übersetzt, jedoch nachfolgend manuell von einem erfahrenen Übersetzer korrigiert, wobei Fehler der MÜ berichtigt werden.

Menschliche Übersetzung Die bestmögliche Qualität einer Übersetzung erreicht man durch einen professionellen Übersetzer mit langjähriger Erfahrung. Obwohl diese Option naturgemäß langsamer und teurer als die vorgenannten beiden ist, ermöglicht die vorliegende Implementation und Workflow durch ausgefeilte Automatismen die schnellste und kostengünstigste Übersetzung, welche man im Internet vorfinden kann.

Es gibt sechs Qualitätsstufen der maschinellen Übersetzung, wobei nicht alle für jedes Sprachpaar vorliegen.

Stufe 0 Dies ist eine sehr grundlegende Form - kaum MÜ. Es handelt sich dabei um eine einfache Wort-für-Wort Übersetzung ohne morphosyntaktische Analyse oder Disambiguierung. Diese Methode ist sehr kostengünstig und der Einsatzbereich für die resultierenden Übersetzungen ist die Möglichkeit den Textinhalt des übersetzten Textes zu verstehen.

Stufe 1 Dieser Modus fügt dem Übersetzungsprozess morphosyntaktische Analyse wie auch Disambiguierung auf Grundlage statistischer Daten der Worthäufigkeit. Es gibt jedoch keine syntaktische Umordnung in der Zielsprache. Diese Methode ist kosteneffizient und das Ergebnis sollte mehr als ausreichend sein um die Bedeutung des Originaltextes zu erfassen. Mindestens diese Qualitätsstufe sollte verwendet werden, wenn eine manuelle Nacheditierung gewünscht wird.

Stufe 2 Fügt dem Übersetzungsprozess die korrekte syntaktische Umordnung in der Zielsprache bei, sowie einen vorausschauende n-Wort Analyse. Die Übersetzungsqualität ist vergleichbar mit derjenigen der meisten modernen MÜ Systeme und kann in allen Fällen verwendet werden, in denen der Benutzer eine zeitgemäße MÜ Qualität akzeptieren kann.

Stufe 3 Fügt elaborierte Methoden zur Disambiguierung sowie die Verarbeitung prozeduraler Wörter hinzu. In regulären Texten ist die Übersetzungsqualität etwas besser als diejenige der Stufe 2 und die resultierenden Texte können sehr gut in Intranets verwendet werden, in denen Informationen in mehreren Sprachen zu geringsten Kosten bereitgestellt werden müssen.

Stufe 4 Fügt semantische Inferenz, NER, sowie Anaphora Resolution hinzu. Wem all diese Begriffe nichts sagen, der sei versichert, daß es sich hierbei um den aktuellen Forschungsstand heutiger MÜ Technologien handelt und folglich bislang unübertroffene Ergebnisse liefert. Die Übersetzungen erreichen publizierbare Qualität für normale Texte und erfordern wenig manuellen Eingriff zur Perfektion.

Stufe 5 Das System durchsucht proaktiv das Internet als Ressource nach relevanten Texten oder Wörterbüchern zusätzlich zu allen lokalen Features um gute Übersetzungen auch für nicht abgedeckte Gebiete zu liefern. Das System extrahiert Informationen aus diesen Quellen und lernt die relevanten Fakten aus den entsprechenden Bereichen, wodurch es die eigene Wissensbasis erweitert.

Warum garantiert nun dieses Framework selbst bei menschlichem Eingriff die schnellste und kostengünstigste Übersetzung wie oben erwähnt? Erreicht wird dies durch einen hohen Automatisierungsgrad unter Einbeziehung der zur Verfügung stehenden sprachverarbeitenden Technologien. So ist z.B. für die Validierung einer menschlichen Übersetzung kein weiterer manueller Eingriff/Lektor mehr notwendig. Bild 13 zeigt eine Gesamtübersicht des Validierungsprozesses für Übersetzungen.

4 Ökonomische Aspekte

4.1 Freie und kommerzielle Nutzung

Das Portal soll eine ausgewogene Balance an frei verfügbaren sowie Bezahltdiensten bieten. Um dies zu gewährleisten wurde ein Konzept entwickelt, welches mit Kosten-“Einheiten” operiert, wobei verschiedene Funktionalitäten auf dem Portal eine entsprechende Anzahl

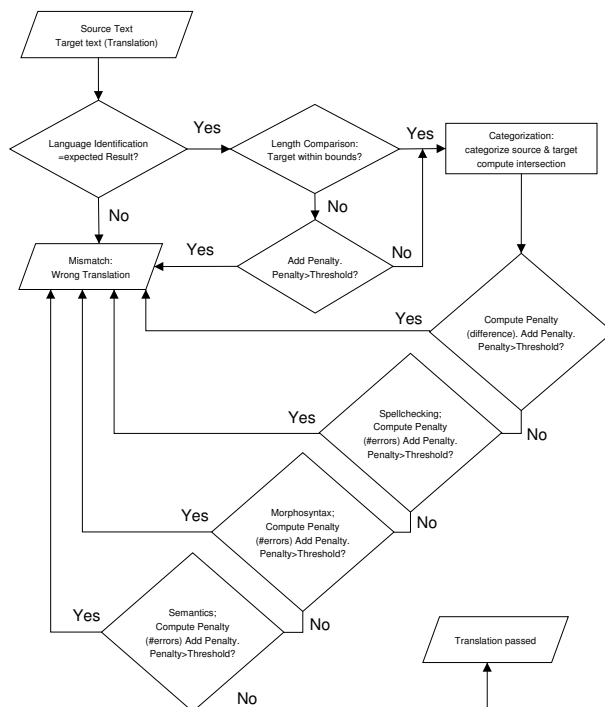


Figure 13: Gesamtübersicht Validierungsprozess Übersetzungen

Einheiten kosten. Darüber hinaus erhält prinzipiell jeder User ein sog. Freikontingent an Einheiten, welche die kostenlose Nutzung der Portaldienste ermöglichen. Eine detaillierte Übersicht des Prinzips sowie Funktionsweise des Freikontingents findet man unter: http://nlp.petamem.com/help.cgi?tid=11&new_site_lang=de&inline=1 (<http://tinyurl.com/6zawf>)

4.1.1 Mitgliedschaft und Freikontingent

Von zentraler Bedeutung ist hierbei die zyklische Erneuerung des Freikontingents, so daß Gelegenheitsnutzer diese Dienste gänzlich kostenfrei nutzen können, wobei professionelle Nutzer (Übersetzer, Lektoren etc.) oder Firmen eher eine Form der Mitgliedschaft wählen sollten. Die Unterschiede zwischen den einzelnen Formen der Mitgliedschaft sind detailliert unter http://nlp.petamem.com/help.cgi?tid=2&new_site_lang=de&inline=1 (<http://tinyurl.com/5zvqy>) aufgeführt. An dieser Stelle sei nur der Übersicht halber die grobe Kategorisierung angegeben:

Guest Gastaccount, kostenfrei, keine Registrierung notwendig, mit dem geringsten Freikontingent, gebunden an eine spezifische IP Adresse.

Registered Benutzer nach erfolgter Registrierung, kostenfrei, höheres Freikontingent, individuell dem Benutzer zugeteilt.

Advanced bezahlte Mitgliedschaft, höheres Freikontingent, zusätzliche Funktionalität, geringere Unit-Kosten

Premium Analog zu “Advanced”, mit weiter verbesserten und umfangreicheren Diensten.

4.2 “Use more, pay less”

Die verschiedenen Formen der Mitgliedschaft sollen natürlich für verschiedene Nutzungsgrade vom Gelegenheitsnutzer bis hin zum “Power-User” ausgelegt sein. Hierbei ist in erster Linie wichtig, für diese Benutzer die richtige Kostenstruktur zu finden. Obwohl die Tendenz dieser Kostenstruktur klar ist - Mehrnutzung soll mit niedrigeren Kosten verbunden sein (eine Art Mengenrabatt), sind die konkreten Parameter für eine ausgewogene Kostenstruktur nicht im Voraus bekannt.

Alle Benutzerdaten sowie die Preislisten (in diesem Fall für Serviceleistungen des Portals) werden von dem CRM/WAWI System (siehe Bild 1) verwaltet und bieten einen Parameter-Raum, bei dem - abhängig von Typ der Mitgliedschaft - Kosten pro Einheit, Größe des Freikontingents bis hin zu einzelnen Benutzern festgelegt werden können.

Das Portal kann darüber hinaus auch aufgrund der Herkunft des Benutzers andere Preislisten auswählen (z.B. um Benutzern aus Pakistan andere Preise anzubieten als Benutzern aus den USA¹³).

4.3 Firmen/Institutionen vs. private Nutzer

Firmen haben häufig andere Anforderungen an ein derartiges Service-Portal als Privatnutzer. Zum einen soll für erbrachte Leistungen eine Rechnungsstellung erfolgen anstelle eines Prepay Modells, zum anderen sollen die angebotenen Dienste mehreren Mitarbeitern zur Verfügung gestellt werden, wobei diese jedoch aus einem gemeinsamen Pool schöpfen sollen.

Beide Anforderungen kann die Accounting Infrastruktur des Portals erfüllen, indem einerseits negative Kontostände zugelassen werden, andererseits mehrere Benutzer zu einem sog. Sammelaccount zusammengefasst werden wobei aber die Informationen über die Individuelle Nutzung innerhalb dieses Pools natürlich erhalten bleibt.

Größere Unternehmen haben jedoch für gewöhnlich eine eigene Instanz des Portals innerhalb ihres eigenen Intranets.

4.4 Vereinbarungen mit Drittanbietern

Es ist klar, daß unabhängig vom Funktionsumfang der PetaMem-eigenen sprachverarbeitenden Lösungen durch diese alleine niemals das gesamte Spektrum an NLP/NLU Diensten abgedeckt werden kann. Im Grundkonzept des Portals sind daher bereits Mechanismen vorgesehen, welche die Kooperation mit spezialisierten Drittanbietern auf allen Ebenen (technisch, wirtschaftlich und rechtlich) erleichtern sollen.

5 Soziale Aspekte

Ein Portal ist im Grunde genommen eine Service-Plattform für einen Benutzerkreis mit gemeinsamen Interessen. Der Nutzen für die Besucher und damit letztlich der Erfolg eines derartigen Portals hängt in starkem Maße davon ab, wie sehr die dargebotene Funktionalität den Interessen entgegenkommt, bzw. wie schnell/gut Anforderungen und Bedürfnisse der Besucher erfüllt werden.

Ein wesentlicher Faktor hierbei ist Kommunikation. Sowohl zwischen Besuchern und der Administration wie auch zwischen den Besuchern untereinander. Probates Mittel um diese Kommunikation zu fördern sind natürlich Diskussionsforen (Message Boards - MB). Bild 14 zeigt einen Ausschnitt der MB Implementierung auf dem NLP Portal.

¹³gegenwärtig nicht realisiert

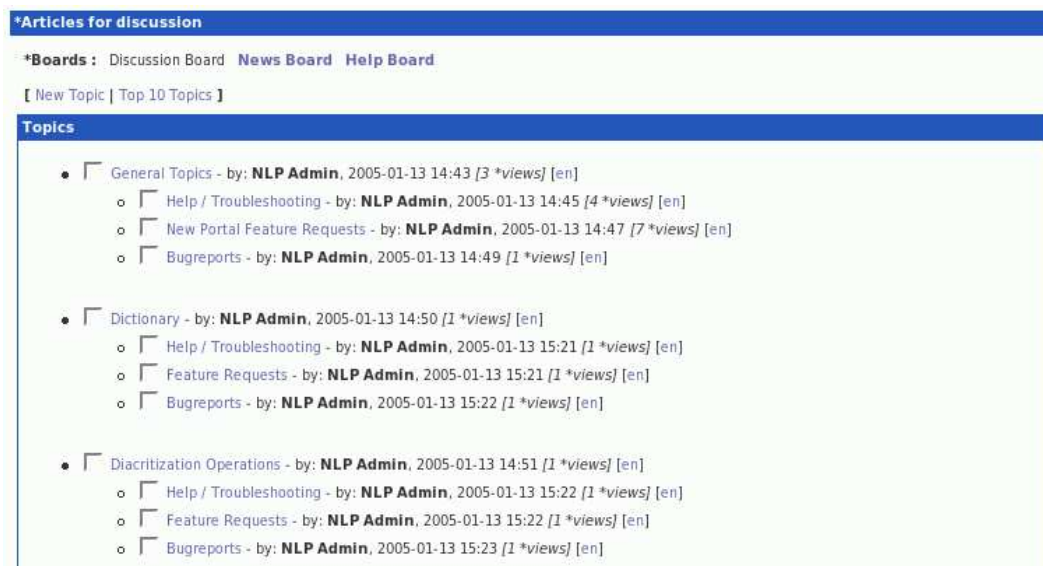


Figure 14: Multilinguales Messageboard.

Hauptmerkmal dieses MB ist die konsequente multilinguale Ausrichtung. Jeder Artikel kann in einer beliebigen¹⁴ Sprache vorliegen - gleichzeitig. Man kann also durchaus auf einen englischen Artikel deutsch antworten, wobei die Übersetzung DE→EN zu einem späteren Zeitpunkt automatisch oder manuell erfolgen kann.

Ein nicht unerheblicher Faktor ist auch die erfolgreiche Einbindung proaktiver Mitglieder in den Entwicklungsprozess des Portals. So können “User Contributions” von der Lokalisierung des Portals, bis hin zur Implementierung eigener NLP Module einfach und schnell integriert werden.

6 Kurz- & mittelfristige Pläne

Wie schon zu Beginn erwähnt, zeigt die gegenwärtige Implementierung des Portals lediglich eine Teilfunktionalität der zugrundeliegenden sprachverarbeitenden Komponenten. Wir planen Kurz- und mittelfristig den gegenwärtigen Funktionsumfang erheblich zu erweitern und das Portal somit als eine Art “one-stop-shopping” Punkt für alle Belange der Sprachverarbeitung im Internet zu etablieren. Die geplanten Erweiterungen sind unter anderem:

- Textkategorisierung und Text-Zusammenfassung (Text summarization)
- Weitere Bezahlssysteme wie auch offline-Bezahlmethoden hinzufügen.
- Lokalisierung in weitere Sprachen fertig stellen.
- Email Frontend für alle NLP Funktionalitäten
- Priorisierung der NLP Anfragen für höhere Benutzerklassen.
- Umfangreiche Fun-Sektion (vgl. Abschnitt 3.2)

¹⁴d.h. vom Portal unterstützten

7 Verweise

NLP Portal Homepage	nlp.petamem.com
Mason Homepage	www.masonhq.com
MT/HT Comparison	www.petamem.com/t_n_p/aslib_tc26.pdf
Lingua::Num2Word	search.cpan.org/~rvasicek/Lingua-Num2Word-0.07/Num2Word.pm
Mosix	www.mosix.org
OpenMosix	openmosix.sourceforge.net

BIO

Richard Jelinek is founder of the PetaMem Group and managing director of its czech and german subsidiaries. Since 1999, Mr. Jelinek was managing director of the czech subsidiary of the SuSE Linux AG, SuSE CR, s.r.o. 1997 master thesis "Qualitative reasoning in a geographical database" at the "Lehrstuhl für Künstliche Intelligenz, Friedrich Alexander Universität" Mr. Jelinek was born in 1970 in Prague/Czech Rep., emigrated 1979 to Germany, and lives since 2000 in Prague again. Main interest: NLP/NLU and its commercialization.